Failure to Reject the *p*-value is Not the Same as Accepting it
The Development, Validation, and Administration of the *KPVMI* Instrument

Rachel Elizabeth Keller

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Curriculum & Instruction

C. Ulrich, Committee Co-Chair
G. Skaggs, Committee Co-Chair
A. Driscoll
E. Johnson

February 27, 2019
Blacksburg, VA

Keywords: *p*-values, research methods, statistical knowledge

Failure to Reject the *p*-value is Not the Same as Accepting it
The Development, Validation, and Administration of the *KPVMI* Instrument

Rachel Elizabeth Keller

ABSTRACT

The purpose of this study was to investigate on a national scale the baseline level of *p*-value fluency of future researchers (i.e., doctoral students). To that end, two research questions were investigated. The first research question, *Can a sufficiently reliable and valid measure of p-value misinterpretations (in a research context) be constructed?,* was addressed via the development and validation of the Keller *P*-value Misinterpretation Inventory instrument (*KPVMI*). An iterative process of expert review, pilot testing, and field testing resulted in an adequately reliable measure (*Alpha* = .8030) of *p*-value fluency as assessed across 18 misinterpretations and 2 process levels as well as an independently validated sub-measure of *p*-value fluency *in context* as assessed across 18 misinterpretations (*Alpha* = .8298). The second research question, *What do the results of the KPVMI administration tell us about the current level of p-value fluency among doctoral students nationally?*, was addressed via analysis of a subset of the field test data (*n* = 147) with respect to performance on the subset of items considered sufficiently validated as developed in Phases I-III (KPVMI-1). The median score was 10/18 items answered correctly indicating that *future* researchers on the aggregate struggle to properly interpret and report *p*-values in context; furthermore, there was insufficient evidence to indicate training and experience are positively correlated with performance. These results aligned with the extant literature regarding the *p*-value misinterpretations of *practicing* researchers.

Failure to Reject the *p*-value is Not the Same as Accepting it
The Development, Validation, and Administration of the *KPVMI* Instrument

Rachel Elizabeth Keller

GENERAL AUDIENCE ABSTRACT

The purpose of this study was to investigate on a national scale the baseline level of *p*-value fluency of future researchers (i.e., doctoral students). To that end, two research questions were investigated. The first research question, *Can a sufficiently reliable and valid measure of p-value misinterpretations (in a research context) be constructed?,* was addressed via the development and validation of the Keller *P*-value Misinterpretation Inventory instrument (*KPVMI*). The second research question, *What do the results of the KPVMI administration tell us about the current level of p-value fluency among doctoral students nationally?,* was addressed via analysis of a subset of the field test data (n = 147) with respect to performance on the subset of items considered sufficiently validated as developed in Phases I-III (KPVMI-1). Results indicate that *future* researchers on the aggregate struggle to properly interpret and report *p*-values in context; furthermore, there was insufficient evidence to indicate training and experience are positively correlated with performance. These results aligned with the extant literature regarding the *p*-value misinterpretations of *practicing* researchers.

Dedication

I dedicate this book to my children – Bud Bud, Mad Dog, Thumbelina, and Sparky – as if somehow that makes up for five years of ignoring them.

Acknowledgements

I would like to acknowledge Dr. Scott Inch without whom I never would have pursued a PhD in the first place and Dr. Robert Gates without whom I never would have had the opportunity.

I would like to acknowledge Dr. Herman Senter without whom I never would have abandoned my stubborn (and unfounded) aversion to statistics and data analysis.

I would like to acknowledge Dr. Lenny Jones without whom I never would have believed I could complete a PhD nor would have had the second chance (you were right about Johns Hopkins…).

I would like to acknowledge Dr. Catherine Ulrich for her support in securing my position in my current program without which I never would have been able to pursue the plan of study my research required. I would also like to acknowledge Katy for agreeing to co-chair my committee, for without her, there would have been no dissertation at all.

I would like to acknowledge Dr. William Woodall without whom I never would have known about the *p*-value ban and thus this entire research agenda would fail to exist.

I would like to acknowledge Dr. Gary Skaggs for his contributions to my methodological training, both in and out of the classroom, and for his professional expertise on this project – both of which elevated the quality of this research. I would also like to acknowledge Gary for his patience through all those meetings and all those questions!

I would like to acknowledge Dr. Anne Driscoll for her invaluable support throughout this ordeal and for her statistical expertise.

I would like to acknowledge Dr. Estrella Johnson for EVERYTHING. When I say that I would not have survived this process without her, I say it without hyperbole.

Finally, I would like to acknowledge the participants in this study without whom none of this would have been possible.

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction

It was observed by Ronald Fisher in 1922 that "the study of statistics, in its theoretical aspects, [had] fallen into prolonged neglect" (p. 310). These words appear a bit pessimistic considering statistical inference was in a period of relative infancy; unfortunately, nearly a century of progress later and there is reason to believe the state of the discipline renders those words just as valid today as when first expressed. Growing concern over the ability of practicing researchers to appropriately interpret and report statistical results has compelled journal editors and professional societies alike to take action in an effort to address and ameliorate the situation. In an effort to investigate whether the renewed attention to this issue has initiated a cycle of self-correction amongst the research community, the purpose of this research is to determine the extent to which *future* researchers are positioned to perpetuate this damaging pattern.

1.1 Background – Statistical Research in Neglect

New disciplines are always subject to criticism in their emergence. As observed by Jerzy Neyman, "the first efforts in one direction contain but rarely the last word to be said" (Fisher, 1935, p. 73). Statistical inference was no exception to this rule. The preeminent scholars of the day (notably, K. Perason, R. Fisher, J. Neyman, E. Pearson) waged a fervent, often contentious decades' long debate concerning issues such as the logic of inductive inference, the role (and validity) of inverse probability, and the appropriate means for testing hypotheses, among others. As the discipline matured through resolution of these foundational conflicts, the procedures concerning statistical inference evolved, eventually reaching asymptotic stability.

Somewhere along the way (at least by 1960 as observed by Rozeboom), this stability was supplanted with relative complacency in which rigorous theoretical study and well-reasoned argumentation was replaced with blind allegiance to ritualistic procedures – regardless of, and often without, sound rationale for said procedures. Rozeboom explains:

> To the experimental scientist…statistical inference is a research instrument, a processing device by which unwieldy masses of raw data may be refined into a product more suitable for assimilation into the corpus of science, and in this lies both strength and weakness. It is strength in that, as an ultimate *consumer* of statistical methods, the experimentalist is in position to demand that the techniques made available to him confirm to his actual needs. But it is also weakness in that, in his need for the tools constructed by a highly technical formal discipline, the experimentalist, who has specialized along other lines, seldom feels competent to extend criticisms or even comments; he is much more likely to make unquestioning application of procedures more or less by rote from persons assumed to be more knowledgeable of statistics than he…Further, since behaviors once exercised tend to crystallize into habits and eventually traditions, it should come as no surprise to find that the tribal rituals for data-processing passed along in graduate courses in experimental method should contain elements justified more by custom than by reason. (p. 416)

To be fair, those words were published in 1960; however, like Fisher's, they also ring true today. To that point, in 2014, George Cobb, posed these questions (and answered them himself) to an American Statistical Association (ASA) discussion forum (Wasserstein & Lazar, 2016):

> Q: Why do so many colleges and grad schools teach *p* = .05?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use *p* = .05?
> A: Because that's what they were taught in college or grad school. (p. 1)

P-values, and their uneasy relationship with hypothesis testing, persist because they are "part of the vocabulary of research" (Goodman, 2008, p.138). These 'laws', "handed down to us by ourselves, through the methodology we adopt" (Nuzzo, p. 150), perpetuate because researchers allow them to – evidence of misuse notwithstanding.

Issues with *p*-value misinterpretation and misuse of significance testing is a complaint so unoriginal that even back in 1966, to complain about it was likened to that of "assuming the role of the child who pointed out that the emperor was really outfitted in his underwear" (Bakan, 1966, p. 423). Despite the fact that statisticians have generally been in agreement for decades now about the deplorable state of statistical inference in scientific research, i.e., "statistical folkways of a more primitive past continue to dominate the local scene" (Rozeboom, 1960, p. 417); the mentality of "We teach it because it's what we do; we do it because it's what we teach" (Wasserstein & Lazar, 2016, p.1) pervades, and borrowing one of Fisher's analogies, acts like an "impenetrable jungle that [arrests] progress towards precision of statistical concepts" (1922, p. 311).

Modern statisticians, like Fisher before them, are aware of the consequences that decades of "neglect of theoretical study of statistics" have imposed on research quality, yet reform remains elusive. As observed by John Campbell (1982), "It is almost impossible to drag authors away from their *p*-values, and the more zeroes after the decimal point, the harder people cling to them" (as cited in Nuzzo, p. 152). Neyman would argue that this is so because "any critical review of basic ideas is always unpleasant and difficult to be accepted by those who are perhaps too much attached and accustomed to the ideas" (Fisher, 1935, p. 73).

The fact that Neyman, Rozeboom, and Cobb all make the same point, despite decades of separation (1935, 1960, and 2014, respectively), is distressing. The fact that reform remains elusive, that the issues "continue to continue [to the extent] that the advance of science has been seriously impeded", in Jacob Cohen's opinion (1994, p. 997), stem from the fact that most statisticians, despite being aware of the rampant misuse, resign themselves to "mostly grouse" in lieu of real action.

While there were some efforts in this direction, the lack of serious action is likely not attributable to apathy, but instead a conditioned response. Articles and books, by Cohen and others (e.g. Goodman, 2008; Meehl, 1967; Gigerenzer, 1993; Falk & Greenbaum, 1995; Morrison & Henkel, 1970) have been published on the subject and clever acronyms have even been assigned to reflect the situation, i.e., Statistical Hypothesis Inference Testing – SHIT (Cohen, 1994) ; however, for every article published warning of the insidious nature of the dubious yet ubiquitous Null Hypothesis Significance Testing (NHST) procedure, another article defending its utility popped up in rebuttal with a meteoric rise of such articles occurring in the 1990s (Robinson & Wainer, 2001). Unable to resolve this impasse, researchers continued submitting, and editors continued accepting, manuscripts rife with methodological flaws despite the fact that the majority of those involved were complicit in their knowledge of "science's dirtiest secret" (Siegfried, 2010, p. 26). As Cohen (1994) explains: teachers, consultants, and authors [have acted] as perpetrators "responsible for the ritualization of null hypothesis significance testing…to the point of meaninglessness and beyond" (p. 997). Although to be fair, there were a couple notable attempts at real action.

The first of these came in 1998, when Kenneth Rothman, founder and editor of *Epidemiology*, penned an article intended to assist prospective authors in the submission process to his journal. In his statement, he offered advice on a myriad of topics, some of which included: grammar, figures and graphs, brevity, and authorship. Buried in the fourth paragraph in the section on data selection and analysis, there was a section essentially banning the use of *p*-values and NHST.

> When writing for **Epidemiology**, you can also *enhance your prospects if you omit tests of statistical significance*. Despite a widespread belief that many journals require significance tests for publication, the Uniform Requirements for Manuscripts Submitted to Biomedical Journals[3] discourages them, and every worthwhile journal will accept papers that omit them entirely. In **Epidemiology**, *we do not publish them at all*. Not only do we eschew publishing claims of the presence or absence of statistical significance, we discourage the use of this type of thinking in the data analysis, such as in the use of stepwise regression. We also would like to see the interpretation of a study based not on statistical significance, or lack of it, for one or more study variables, but rather on careful quantitative consideration of the data in light of competing explanations for the findings. For example, we prefer a researcher to consider whether the magnitude of an estimated

> effect could be readily explained by uncontrolled confounding or selection biases, rather than simply to offer the uninspired interpretation that the estimated effect is "significant," as if neither chance nor bias could then account for the findings.
>
> Many data analysts appear to remain oblivious to the qualitative nature of significance testing. Although calculations based on mountains of valuable quantitative information may go into it, statistical significance is itself only a dichotomous indicator. As it has only two values, "significant" or "not significant," it cannot convey much useful information. Even worse, those two values often signal just the wrong interpretation. These misleading signals occur when a trivial effect is found to be "significant," as often happens in large studies, or when a strong relation is found "nonsignificant," as often happens in small studies. *P*-values, being more quantitative, are preferable to statements about statistical significance tests, and *we do publish P-values on occasion*. We do not publish them as an inequality, such as $P < 0.05$, but as a number, such as $P = 0.13$. By giving the actual value, one avoids the problem of dichotomizing the continuous *P*-value into a two-valued measure. Nevertheless, *P*-values still confound effect size with study size,[4] the two components of estimation that we believe need to be reported separately. Therefore, *we prefer that P-values be omitted altogether*, provided that point and interval estimates, or some equivalent, are available. (p. 334)

This radical notion, while a valiant attempt to change practice, was underscored by its inconspicuous placement and lack of prominent typeface (emphasis added); so much so that it appears to have gone largely unnoticed by those whose research lay outside the biomedical community. In 2001, the editorship of *Epidemiology* changed hands, and the new editors – having been inundated with letters and requests – published a special statement essentially reversing Rothman’s position asserting that he had in fact never banned them in the first place.

> Does all this mean a change in Epidemiology’s policy on *P*-values? It may be no more than a change in perception. We will not ban *P*-values. But neither did Rothman. He called for caution, and we do the same. The question is not whether the *P*-value is intrinsically bad, but whether it too easily substitutes for the thoughtful integration of evidence and reasoning. Given the *P*-value’s blighted history, researchers who would employ the *P*-value take on a particularly heavy burden to do so wisely. (p. 286)

For over a decade, the statistical community at large was primarily silent on this issue – the interpretation of *silent* being that while grousing did continue, no effectual action was taken.

That is until 2015, when David Trafimow and Michael Marks “went rogue” (Dozo, 2015) and, in what was arguably misreported by *Scientific American* as “the first such move ever by a scientific journal” (Nuzzo, 2015), banned *p*-values from *The Journal of Basic and Applied Social Psychology*, and did so with significant fanfare and conviction so that there would be no misinterpretation as had clouded Rothman’s statement.

> The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would not be required to perform it (Trafimow, 2015). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP. (p.1)

The editors presented clarifying information in a Q&A format, and ended the article with this final thought:

> Some might view the NHSTP ban as indicating that it will be easier to publish in BASP, or that less rigorous manuscripts will be acceptable. This is not so. On the contrary, we believe that the $p < .05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research. We hope and anticipate that banning the NHSTP will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking. The NHSTP has dominated psychology for decades; we hope that by instituting the first NHSTP ban, we demonstrate that psychology does not need the crutch of the NHSTP, and that other journals may follow suit. (p. 2)

Whether or not other journals will follow suit remains to be seen; however, the firestorm generated with this editorial was significant. Researchers across disciplines rushed to weigh in with varying levels of support and disdain for the journal’s policy. The debate reached such epic proportions as to elicit reactions from the president of the Royal Statistical Society and the board of directors of the American Statistical Association (ASA). In response to all the attention, Trafimow defended his position thusly: “If scientists are depending on a process that’s blatantly invalid, we should get rid of it…I'd rather not have any inferential statistics at all than ones that we know aren't valid" (Woolston, 2015).

Of those weighing in on the debate, the majority of those celebrating the decision used similar rhetoric attacking the reliability and objectivity of *p*-values, their suitability in hypothesis testing, and the inability of the majority of scientists to interpret them properly. One of the more impassioned examples asks "Why should we test improbable and irrelevant null hypotheses with a chronically misunderstood and abused method with little or no scientific value that has several large, detrimental effects even if used correctly (which it rarely is)?" (Karlsson, 2014).

While there appears to be consensus that *p*-values are "misinterpreted, misused, and abused both by naïve analysts and statisticians" (Irizarry, Peng, & Leek, 2014), those condemning the ban argue that quitting cold turkey is like throwing the baby out with the bathwater. Peter Diggle, president of the Royal Statistical Society, said "The RSS welcomes and shares the BASP editors' concerns that inferential statistical methods are open to mis-use [sic] and mis-interpretation [sic], but does not feel that a blanket ban on any particular inferential method is the most constructive response" (Editor, 2015). The ASA shared his concern that conflation of the invalidity of a particular statistical technique with researchers' inability to properly employ it is not only foolish but potentially dangerous. In their formal statement, the organization warned of the "negative consequences" of such a ban and suggested that "the proper use of inferential methods needs to be analyzed and debated in the larger research community" (Editor, 2015). This dissertation aims to be a part of that conversation.

1.2 Study Rationale

It is not the objective of this dissertation to take a position on the debate over the appropriateness of the *p*-value ban, but rather to justify its own existence in that a research study of this nature is a logical consequence of such a measure. This research is timely, not because *p*-value misinterpretation is in any way a new issue, but because the ban and the backlash it generated, has elevated the topic to an international conversation. The research literature on *p*-value misuse dates back several decades; however, the current level of notoriety extends beyond academia to the popular media and has even garnered enough attention as to warrant its own Wikipedia page (https://en.wikipedia.org/wiki/Misunderstandings_of_p-values).

Even the American Statistical Association has felt compelled to join the conversation. In the 175 years of its existence, the ASA had *never* taken an official position on specific matters of statistical practice, but yet concern about "issues of reproducibility and replicability of scientific conclusions" (Wasserstein & Lazar, 2016, p. 2) proved too important to ignore. A panel of two dozen experts was assembled and given the task of drafting the statement – a process hampered by controversy that took over a year to complete. The statement, released in early 2016, was aimed at researchers and practitioners with the intention of "provid[ing] the community a service" (p.3); in other words, to help the non-statistician apply statistics in research.

While misinterpretation of statistical concepts is not limited to *p*-values or the current generation, the reason the ASA felt compelled to intervene now is likely due to the surge in significance testing by non-statisticians. As early as 1940, William Ogburn forecasted this future ubiquity:

> "In the early days when statistical literacy was low, those who could read and write this strange new language were set off and apart from the others. They were labeled statisticians. But now almost any social scientist can compute a correlation coefficient, and can read and write the statistical language to some extent." (p. 260)

Across disciplines, "*p*-values [are] used to back claims for the discovery of real effects amid noisy data" (Matthews, 2017, p.38); however, the validity of such findings is being called into question as it is now widely accepted that the majority of scientists can neither properly explain nor aptly interpret *p*-values in the context of hypothesis testing (e.g. Goodman, 2008, Badenes-Ribera, et al., 2015).

The point of statistics, as described by Lambdin (2012), is to “clear the fog and help us identify patterns and relationships in cumbersome data that the naked eye cannot detect unaided”; however, it is quite frequently the case that “statistics instead serves – indeed, is all too often intentionally employed – to smokescreen what the naked eye does indeed see unaided: that there’s nothing there” (p. 84). When researchers conflate flukes with insights (Matthews, 2017), potential collateral damage can be inflicted upon unsuspecting populations when policy changes are implemented on the basis of research that is un-replicable.

“Poor reporting of the statistical analysis does not necessarily mean that the conclusions will be wrong”, Norstrom (2015) explains; however, a key consequence is that “readers will not be able to critically assess whether the statistical analyses provide reliable results or whether the conclusions drawn by the author are valid”. Incorrect conclusions have the potential to have profound consequences, especially in education. Politicians and school executives look to research, much of which is faulty, to make policy decisions. This is especially true in the wake of the No Child Left Behind Act of 2001 because the Act “placed great emphasis on using scientific research to determine what works best in our schools” (Suter, 2012, p. 4). NCLB, coupled with the U.S. Department of Education’s Institute of Education Sciences influence in shifting educator’s focus to utilizing research to identify best practices, “highlights the value of empirical, systematic, rigorous, objective [research] to obtain valid knowledge about teaching and learning” (pg. 4).

An increased emphasis on research-based practice and policy is prudent, but does rely heavily on the quality of the research. Weintraub (Deubel, 2008) explains that when education decisions are based on misinterpretation of research results, the policy changes "have not necessarily led to increases in student achievement". He cites, among his examples, increased spending, charter schools, supplemental education services, and decisions aimed at increasing accountability. Failure to increase achievement is problematic from a financial standpoint in that these expenditures have not served their purpose and in being exhausted cannot be reallocated to better use. Gerald Bracey, in his indictment of the *A Nation at Risk Report*, illustrates a situation with even greater consequences. The 1983 report claimed a "rising tide of mediocrity" and was referred to as 'Paper Sputnik' for all the shame and panic it bestowed upon the country (as described by Suter, 2012); however, Bracey claims this was a "manufactured crisis". In 2006, he returned to the data to investigate and verify the original claims and what he found was a "golden treasury of selected, spun, distorted, and even manufactured statistics" (2008, p. 63). That is not to say that *p*-value misinterpretation was solely to blame, but this report and the sweeping reform efforts that stemmed from it, is a "prime example of how research in education can have a huge and lasting impact" (Suter, 2012, p. 37).

There is no doubt that public policy decisions based on tenuous research findings have the potential to be harmful; yet, it is important to also acknowledge the damaging effects that publishing inferior research have on the researchers and the field itself. It is conceivable that one might view the *p*-value ban by the *BASP* editors as overreach in that it represents a confiscation of methodological tools from researcher's toolboxes owing to a lack of confidence that they can be implemented responsibly and properly. Not wishing to dismiss that viewpoint, but to augment it, one might view the ban as an important precautionary intervention to protect the research community from inflicting harm on itself in much the same way a parent removes scissors from a small child to prevent injury.

When theoretically questionable procedures, like NHST, become the norm, and reporting of results rife with statements reflecting *p*-value misconceptions become the rule rather than the exception, it becomes increasingly difficult for editors and reviewers to filter proper and improper application of these statistical tools in submissions. Novice authors and graduate students are likely to mimic the reporting style of the papers in their target journals in the hopes of increasing their odds of acceptance. Thus the new generation replicates the mistakes and ensures the tradition of inferior reporting continues. When the same statements (*highly* significant or *nearly* significant) and practices ($p < .05$ in lieu of $p = .024$) are seen repeatedly, they cease to be regarded as errors for they come to symbolize just the way things are done.

This is somewhat of an insidious problem because, for the most part, it seems that researchers commit these errors unconsciously. While exceptions may exist, it is doubtful that the majority of professionals who conduct and report research do so with the intent to mislead or misrepresent; however, in misreporting results as a consequence of misconceptions concerning statistical inference, this happens despite their virtuous intentions. Published papers, serving as training documents for graduate students, unwittingly teach quantitative methodology to the next generation of researchers in a manner that is inconsistent with statistical theory.

This is a problem that is confounded further when considering the necessity of citing previous related work in one's literature review. Sarewitz (2016, p.147) has labeled this problem one of "a destructive feedback between the production of poor-quality science, the responsibility to cite previous work, and the compulsion to publish"; a problem he contends is "likely to be worse in policy-relevant fields such as education, in which the science is often uncertain and the societal stakes can be high". Many researchers, despite the awareness of the prevalence of unreliable research studies, expect that the scientific literature will self-correct over time;

however, Picho and Artino (2016, p.483) argue that this is not the case. They cite the *file drawer effect* (unpublished studies with negative outcomes) and the underappreciation for replication as evidence that “self-correction of faulty results may be the exception, not the rule”. Repeated citations of unreliable papers serve to legitimize them and obfuscate the field’s ability to self-correct.

One can imagine how replicate citations could replace replicate outcomes in confirmation of findings. As the popularity of a citation rises, so does the apparent credibility of the result which in turn begets more citations. In this case, researchers are less likely to cross-reference with corroborating studies and/or attempt a replication study, but this is a dangerous choice because most research, as it turns out, is unreplicable. According to the results of the Reproducibility Project, nearly 2/3 of 100 replication attempts of studies published in highly respected journals failed (Baker, 2015). The so-called “replication crisis”, “in which widely cited research claims fade away on reinvestigation” (Matthews, 2017, p.38) is undeniably linked to the unreliability of *p*-values.

A blanket ban on such an important statistical tool might seem rather drastic; however, perhaps the editors of BASP felt that desperate times called for desperate measures. Statisticians, Matthews (2017) observes, are “clearly disturbed by what passes for inference in most scientific disciplines” (p.40).   While oxymoronic to construe inference as a thoughtless act, Lambdin (2012) argues that significance testing has become a “mindless ritual” that researchers apply “with little appreciation of its history and virtually no understanding of its actual meaning” and “despite this alarming dearth of statistical insight, [hold it] up as the hallmark of confirmatory evidence” (p.68). Together, the *p*-value ban and the ASA’s p-value statement could cause scientists to rethink their stance on hypothesis testing, *p*-values, and inference more broadly. The ASA is operating under the expectation, or at least the hope, that what will follow "is a broad discussion across the scientific community that leads to a more nuanced approach to interpreting, communicating, and using the results of statistical methods in research" (Wasserstein & Lazar, 2016, p. 2).

Two years removed from the ban and only one from the ASA's statement, the *p*-value conversation is not only still ongoing, but is current enough to warrant new participants in the discussion. In brief, the release of the ASA statement invites investigation as to whether members of the research community understand and can(are) apply(ing) the statement's principles in practice and the development of such an instrument to measure that capability. This research aims to contribute to that discussion with the development of an instrument designed to measure the *p*-value fluency of future researchers and by providing preliminary results therein based on its administration in a national sample.

1.3 Study Description

The primary purpose of this research study is to determine the extent to which doctoral students, i.e., future researchers, struggle with interpreting and reporting *p*-values in the context of independent research and peer review. To that end, two research questions were investigated. The first research question, *Can a sufficiently reliable and valid measure of p-value misinterpretations (in a research context) be constructed*?, was addressed via the development and validation of the *Keller P-value Misinterpretation Inventory* instrument (KPVMI-1). This endeavor occurred across three phases.

During Phase I, a test blueprint was generated for the instrument based on identified misinterpretations from the literature. An initial item pool, PV0, was generated and subjected to expert review by members of the dissertation committee. Modifications were made in accordance with advice from the statistical advisors and the preliminary version of the instrument, PV1, was created.

During Phase II, the instrument was piloted with Virginia Tech (VT) students (n = 15). Results from cognitive labs (a form of think-aloud interview) and a usability study (n = 5 and n = 10, respectively) informed a revised version of the instrument, PV2.

During Phase III, the instrument was twice field tested: first using students enrolled in a large interdisciplinary graduate research methods course here at Virginia Tech (n = 79), and later in a national sample (n = 207). Instrument validation (to include item analysis and reliability/validity diagnostics) occurred at this stage. A subset of items was identified to serve as the (current) final[1] version of the instrument (PV3).

The second research question, *What do the results of the KPVMI-1 administration tell us about the current level of p-value fluency among doctoral students nationally,* was addressed via analysis of a subset of the field test data (n = 147) with respect to performance on the subset of items considered sufficiently validated as developed in Phases I-III (*KPVMI*-1). Inferences about respondents' *p*-value fluency were drawn from this data, in totality and by subgroup where appropriate.

1.4 Structure of the Dissertation

Chapter 2 is a review of the literature. This section begins with a discussion of the historical evolution of *p*-values and how they came to be used in research, including their relationship with hypothesis testing and how that contributes to their misuse. Following that, existing studies with identified misconceptions are summarized and critiqued. Particular

[1] The *final* version of the instrument in terms of *this* research study. As discussed in Chapter 5, the instrument validation is a work in process and future data collection and instrument revision is planned.

attention is paid to the instruments utilized in previous research so as to highlight the need for the assessment being designed in the present study.

Chapter 3 describes and justifies the methods employed to develop and validate the KPVMI-1 instrument. Chapter 4 presents the data obtained in pursuit of answering the two research questions in two sections. The first section provides the supporting evidence (expert rater comments, student interview transcript data, item response theory statistics) that was used to inform the evolution of the instrument through its various revisions. The second section summarizes the results of the large-scale administration of the instrument. Chapter 5 discusses doctoral students' *p*-value fluency, its compatibility with the ASA's principles for research, and any programmatic (methods course requirements) and/or curricular (alternate presentation of concepts) modifications indicated by the results. The dissertation closes with a presentation of the study limitations and suggestions for future directions of this research.

Chapter 2 – Literature Review

In this chapter, the literature relevant to the development of the PV instrument is presented. The discussion begins with the historical evolution of *p*-values and how they came to be used in research, including their relationship with hypothesis testing, and how that contributes to their misuse. Following that, identified misconceptions and misunderstandings are described and catalogued. An overview of existing research instruments is provided, including a discussion of their limitations and inability to address the present research questions, in order to justify the need for the development of the PV assessment. This chapter ends with the presentation of the problem statement and the research questions.

2.1 History of the P-value

2.1.1 Fisher's Contribution

The use of *p*-values can be traced back to Laplace in the 1770s and his historic investigation of the disproportionate number of male births, yet the credit for their utilization as a formal research tool lies largely with Ronald Fisher and his 1925 *Statistical Methods for Research Workers* (SMRW) handbook. While Fisher was neither the developer of, nor the first to publish on, hypothesis testing and the constituent role of *p*-values therein, he is credited for his introduction of a somewhat formulaic method for performing such tests and the first widely publicized guidelines concerning their significance.

It is of great interest, albeit little consequence, that the "father of statistics" (see e.g. Savage, 1976; Rao, 1992), both metaphorically and literally (he coined the term 'statistic' in similar fashion to its modern-day interpretation), was in fact an outsider – not a statistician, but a geneticist; a point made rather finely by Leonard Savage (1976): "I occasionally meet geneticists who ask me whether it is true that the great geneticist R.A. Fisher was also an important statistician." It was his work as a scientist that motivated his work as a statistician in that he wished to supply practical experimenters – and incidentally, teachers of statistics – an account of laboratory applications for statistical theory (Fisher, 1935). In Fisher's words taken from the introduction to his first edition: "The prime object of this book is to put into the hands of research workers, and especially of biologists, the means of applying statistical tests accurately to numerical data accumulated in their own laboratories or available in the literature" (p. 16).

Of all the potential ways that theory could be applied practically in the laboratory, it was Fisher's (1925) view that there was no more "pressing need" (p. 211) in statistical examination of a body of data than to test its agreement with a suggested hypothesis, the act of which he described as an alternating process of deductive and inductive reasoning. "A hypothesis is conceived and defined with necessary exactitude; its consequences are deduced by a deductive argument; these consequences are compared with the available observations; if these are completely in accord with the deductions, the hypothesis may stand at any rate until fresh observations are available" (p. 9).

In Fisher's system, hypothesis testing (by our nomenclature), consisted of a single hypothesis based upon a previous sample or a proposed theoretical underlying population distribution and the objective was to analyze the strength of evidence in the current sample against the stated hypothesis. "There was no mention of 'error rates' or hypothesis 'rejection', it was meant to be an evidential tool, to be used flexibly within the context of a given problem" (Goodman, 2008, p.135). Fisher describes this process broadly in SMRW:

> The idea of an infinite **population** distributed in a **frequency distribution** in respect of one or more characters is fundamental to all statistical work. From a limited experience, for example, of individuals of a species, or of the weather of a locality, we may obtain some idea of the infinite hypothetical population from which our sample is drawn, and so of the probable nature of future samples to which our conclusions are to be applied. If a second sample belies this expectation we infer that it is, in the language of statistics,

> drawn from a different population; that the treatment to which the second sample of organisms had been exposed did in fact make a material difference, or that the climate (or the methods of measuring it) had materially altered. Critical tests of this kind may be called tests of significance, and when such tests are available we may discover whether a second sample is or is not significantly different from the first. (p.43, emphasis is in the original)

This excerpt, taken from the third chapter of the book, is the first mention and thus *definition* of tests of significance, although definition is used loosely here as his neither defines nor specifies significance. A few pages later, he attempts to illustrate significance by use of *p*-values in the context of the Normal distribution:

> In practical applications we do not so often want to know the frequency at any distance from the centre [sic] as the total frequency beyond that distance; this is represented by the area of the tail of the curve cut off at any point… in other words, what is the probability that a value so distributed, chosen at random, shall exceed a given deviation. Tables I. and II. have been constructed to show the deviations corresponding to different values of this probability. The rapidity with which the probability falls off as the deviation increases is well shown in these tables. A deviation exceeding the standard deviation occurs about once in three trials. Twice the standard deviation is exceeded only about once in 22 trials, thrice the standard deviation only once in 370 trials, while Table II. shows that to exceed the standard deviation sixfold would need nearly a thousand million trials. The value for which P =·.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion, we should be led to follow up a negative result only once in 22 trials, even if the statistics are the only guide available. Small effects would still escape notice if the data were insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty. Some little confusion is sometimes introduced by the fact that in some cases we wish to know the probability that the deviation, known to be positive, shall exceed an observed value, whereas in other cases the probability required is that a deviation, which is equally frequently positive and negative, shall exceed an observed value; the latter probability is always half the former. (p. 46-48)

In customary Fisher fashion, the mathematics is “ruthlessly omitted” (Savage, p.448) from this text. It had been suggested of Fisher by a contemporary (Greenwood, as quoted in

Fisher, 1935) that he possessed "surely a rare defect in statisticians – an extreme reluctance to bore his readers", a point that manifest itself in his over-anxiousness not to incur the sneer of those who might deem that which he said to be 'obvious' or 'self-evident' (p. 68). While certainly plausible in his technical papers, the reason for this style in his practitioner handbooks was to save the non-mathematician from details he considered irrelevant to his use of the applications – a style not too dissimilar to that of current methodology textbooks. In the preface to the fifth edition of SMRW, he explains this.

> Modern statistics could not have been developed without the elaboration of a system of ideas, logical and mathematical, which, however fascinating in themselves cannot be regarded as a necessary part of the equipment of every research worker…the practical application of general theorems is a different art from their establishment by mathematical proof and one useful to many to whom the other is unnecessary. (p. ix – x)

Despite his tendency to "freely pour out mathematical facts…without even a bow in the direction of demonstration" (Savage, p.448), this text speaks of *p*-values, including the oft-confusing notion of one-sided versus two-sided, in detail but with an almost conversational familiarity and fluency. So casually, in fact, that the word *clearly* is almost conspicuously absent; although the interpretation would have been literal for Fisher and not the modern day usage in which the professor is the only one for which the predicate is *clear*. From this, it is assumed that the formal definition of *p*-values was so generally well-established at this point that even lowly research workers would have been expected to understand enough probability to comprehend this description in spite of its brevity. As a historical reference, Karl Pearson published on using a table of values for *p* in 1900 (p. 160) – a full quarter century before Fisher's text was published.

> Suppose we…desire to ascertain whether an outlying observed set is really anomalous…the [table] will give the value of P, the probability of a system of deviations as great or greater than the outlier in question. For many practical purposes, the rough interpolation which this table affords will enable us to ascertain the general order of probability or improbability of the observed result, and this is usually what we want.

Pearson's article, without mentioning it by name, is referring to the act of testing a sample relative to expectation. Thus, Fisher's contribution lies outside of *p*-values and even the notion of hypothesis testing itself. What distinguished Fisher in this regard is the notion that evaluative judgement of the probability of the observed result could be quantified – and delineated; this important observation (recommendation) being arguably the greatest legacy of his seminal work despite his tendency to bury the lead.

While Fisher's decision to abbreviate the mathematical details was intentional, the resulting awkward, contrived style was possibly an inadvertent consequence, or perhaps it was a habit he never felt compelled to outgrow. Whatever the reason, his texts often presented as a series of illustrative examples in which the key details were hidden in plain sight (as opposed to carefully arranged and highlighted as in modern texts) and subject to extraction by the reader. Fortunately, he often reiterated such ideas with varying degree of clarifying information, although it was left to the reader to string these far-flung references together into one coherent understanding. On the topic of significance, the first edition of SMRW discussed this in several places, for example:

> In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05, and consider that higher values of χ^2 indicate a real discrepancy. (p. 80)
>
> …the chance of exceeding which value is between .01 and .02; if we take P=.05 as the limit of significant deviation, we shall say that in this case the deviations from expectation are clearly significant. (p. 82)

Fisher was at times "aphoristic and cryptic" (Savage, p.448) and reading between the lines was often required. However, in use of the word *significant*, this was not one of those times. Fisher was deliberate and intended for people to take him quite literally – the common language interpretation was something worthy of notice (Goodman, 2008, p. 135) – and he did not expect nor relish the idea that people would ascribe a dichotomous, dictatorial role to the *p*-value. Intended to act as a barometer, Fisher operated under the principle that small *p*-values indicated that future sampling was warranted in order to rule out the possibility that the observed effects were the result of chance. Unlike in the modern approach to significance testing in which sampling is often repeated *until* significance is found, Fisher's approach suggested just the opposite. *Significant* meant "worthy of attention in the form of meriting more experimentation, but not proof in itself" (Goodman, 2008, p.135).

It would appear that Fisher felt this interpretation of significant to be self-evident due to the lack of elaboration in *SMRW*; however, his later publications address this point in finer detail. First, in a 1926 journal article: "Personally, the writer prefers to set a low standard of significance, at the 5 percent point…A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance" (p. 504). And again, in his acclaimed follow-up text to *SMRW, The Design of Experiments*:

> ...we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the "one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*. In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. (reprint of 8th Ed., 1971, p. 13-14)

Fisher's interpretation of significance and of hypothesis testing enjoyed a relatively peaceful reign for a few years. His SMRW textbook was in its second edition before J. Neyman and E. Pearson published on the matter. Although the feud between the two sides would later become public and contentious (incidentally, not Fisher's only or even worst feud), the tone of their initial article did not in hindsight appear to be antagonistic so much as elaborative.

It is often thought that the main difference between the approaches of Fisher and Neyman-Pearson is that "in the latter, the test is carried out at a fixed level, whereas the principal outcome of the former is the statement of a *p*-value that may or may not be followed by a pronouncement concerning significance" (Lehmann, 1993, p.1243). While it is true that Fisher deplored this notion of *fixed significance*, it is somewhat open to interpretation if this is literally what Neyman and Pearson had advocated. It was however a different matter entirely that prompted their initial research.

2.1.2 The Neyman-Pearson Contribution

Significance testing as dictated by Fisher involved calculating the deviation of the test statistic from the expectation under the hypothesis and using the relative probability of such a deviation (*p*) as a measure of the strength of evidence in support of the hypothesis. It was not the threshold of significance per se that bothered Neyman and Pearson, but the lack of justification for the grounds on which the decision was being made. While it would have been typical for Fisher's handbooks like SMRW to omit such details, the origins of his test statistics were apparently conspicuously missing from his more technical publications as well and this is what attracted their attention. Here they describe their motivation: "partly because we shall approach the problem from a somewhat different point of view and partly because we believe that some may have difficulty in following Dr. Fisher's very condensed reasoning we propose to go into this question at some length" (Neyman & Pearson, 1928, p. 268).

If nothing else, their quest for length was successful. In 1928, in an article so extensive it was published in two parts comprising 98 pages, Neyman and Pearson set to the important work of understanding, clarifying, and extending the work of Fisher (and others before him) regarding hypothesis testing:

> In general the method of procedure is to apply certain tests or criteria, the results of which will enable the investigator to decide with a greater or less degree of confidence whether to accept or reject [the hypothesis], or, as is often the case, will show him that further data are required before a decision can be reached…The tests themselves give no final verdict, but as tools help the worker…to form his final decision; one man may prefer to use one method, a second another, and yet in the long run there may be little to choose between the value of their conclusions. What is of chief importance in order that a sound judgment may be formed is that the method adopted, its scope and its limitation, should be clearly understood, and it is because we believe this often not to be the case that it has seemed worth while [sic] to us to discuss the principles involved in some detail.
> (p. 176)

In this paper, questioning the reliability of common sense without outright disparaging it, they attempt to mathematically justify why certain criteria (i.e., test statistics) which they later describe as "appear[ing] often to have been fixed by a happy intuition" (Neyman & Pearson, 1933, p.291) are superior to others in their appropriateness for testing hypotheses. What follows is an incredibly mathematically dense discussion illustrating the various criterion and respective contour systems (the bounds of which demarcate rejection/acceptance of the hypothesis) that can be applied to normal, rectangular, and exponential populations – the result of which is the nontrivial Neyman-Pearson (N-P) lemma that established the likelihood ratio test as the most powerful for a given significance level.

From a modern perspective, this problem of test specification might appear to be on the periphery of the hypothesis testing conversation; however, it is important not to underestimate the magnitude of Neyman and Pearson's contribution. In current practice, statisticians select estimators for the properties they possess (i.e., efficiency, unbiasedness, uniformly most powerful, sufficiency, minimum variance, etc.); however, in the 1920s, these notions remained unproven if they had even appeared at all. In Fisher's case, he explained how to use a *p*-value to determine if the test statistic in hand, calculated from the obtained sample, provided worthy evidence in relation to the proposed hypothesis without justification as to why that test statistic had been selected in the first place and without consideration of how the *p*-value might change if alternative criterion had been employed. Without wishing to take his word for it, Neyman and Pearson set to the task of testing multiple criterion in order to determine the best – and for that determination, an objective measure of valuation was necessary.

In what to them seemed an obvious point of departure, Neyman and Pearson felt it necessary to consider this problem of test specification from the perspective of both known approaches to hypothesis testing: "one to start from the population ∏, and to ask what is the probability that a sample such as ∑ should have been drawn from it, and the other the inverse method of starting from ∑ and seeking the probability that ∏ is the population sampled" (1928, p. 176). This two-prong tactic is what led them to propose the "first" and "second sources of error" (1928, p. 177) and what ultimately led to the methodological drift away from Fisher. However, in the 1928 paper, their ideas were yet emergent and only vaguely resembled the acceptance procedure approach for which they would come to be associated.

> Perhaps the most suggestive method…is to represent ∑ by a point in a hyperspace whose dimensions will depend upon the particular problem considered; and to associate the criteria for acceptance or rejection with a system of contours in this space, so chosen that in moving out from contour to contour [the hypothesis] becomes less and less probable…Setting aside the probability that the sampling has not been random or that the population has changed during its course, ∑ must either have been drawn randomly from ∏ or ∏′, where the latter is some other population which may have any one of an infinite variety of forms differing only slightly or very greatly from ∏. The nature of the problem is such that it is impossible to find criteria from which will distinguish exactly between these two alternatives, and whatever method we adopt two sources of error must arise: (1) Sometimes, when [the hypothesis] is rejected, ∑ will in fact have been drawn from ∏. (2) More often, in accepting [the hypothesis], ∑ will really have been drawn from ∏′. (p. 176-177)

In reading this passage, the notion of ‘acceptance or rejection’ is somewhat misleading to the modern reader in that it likely conjures up images of a rejection region based on a particular significance level of a specific test statistic (i.e., the appropriate shaded area under the distributional curve); however, while acceptance can and does come to embody that interpretation, in this instance Neyman and Pearson were referring to acceptance from the perspective of test specification. Each criterion under analysis would have with it an associated system of contours, subject to confidence levels, and that could vary considerably.

Type I and Type II errors (in modern vernacular) are introduced here for the first time as a natural consequence of the decision-making procedure; the contour system for a given test statistic would have a unique, but mathematically determinable, margin of error and the idea was to accept the criterion that optimized the utility subject to the allowable error. Like Fisher’s early work, the threshold for significance was open to interpretation and Neyman and Pearson make suggestions but do not draw a hard line.

> In the long run of statistical experience the frequency of the first source of error (or in a single instance its probability) can be controlled by choosing as a discriminating contour, one outside which the frequency of occurrence of samples from $\prod$ is very small – say, 5 in 100 or 5 in 1000…The second source of error is more difficult to control, but if wrong judgments cannot be avoided, their seriousness will at any rate be diminished if on the whole [the hypothesis] is wrongly accepted only in cases where the true sampled population, $\prod'$, differs but slightly from $\prod$. It is not of course possible to determine $\prod'$, but making use of some clearly defined conception of probability we may determine a ‘probable’ or ‘likely’ form of it and hence fix the contours so that in moving ‘inwards’ across them the difference between $\prod$ and the population from which it is ‘most likely’ that $\sum$ has been sampled should become less and less…Both these aspects of the problem must be taken into account. Using only the first control…a criterion of any desired degree of stringency could always be found for a sample. But regarding the position from the second point of view, it is seen that there will only be certain systems which are of any value as criteria. (1928, p. 177-178)

In the years that followed their seminal piece, Neyman and Pearson developed the idea of a "rule of behavior to be applied repeatedly…when faced with the same set of alternate hypotheses" (1933b, p.509) that consisted of rejecting the hypothesis "when a specified character, x, of the sample lies within certain critical limits, and accepting it or remaining in doubt in all other cases" (1933a, p.295). Acknowledging that in making a decision, one runs the risk of drawing an incorrect conclusion, they succinctly reiterate the two possible kinds of errors: "If we reject H_0, we may reject it when it is true; if we accept H_0, we may be accepting it when it is false, that is to say, when really some alternative H_t is true" (1933a, p.296); however, they explain that the incidence of such errors is both minimal and, perhaps more importantly, governable.

> But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false" (1933a, p. 291).

Neyman and Pearson applied mathematical theory (i.e., Calculus of Variations) to show how the choice of the critical region for a given test statistic could control and minimize the two sources of error. They were successful – at least for the case of testing a single distribution hypothesis against a simple alternative – in answering their own question of test specification using the Type I and Type II error rates (in modern parlance) as their method of appraisal. Their conclusion was that the "best test" criterion was that which minimized "Error II" subject to a constraint on "Error I"; here, in their own words:

> For testing any given statistical hypothesis, H_0, with regard to alternative H_t, if θ_1 and θ_2 are two possible criteria and if in using them there is the same chance, ε, of rejecting H_0 when it is in fact true, we should choose that one of the two which assures the minimum chance of accepting H_0 when the true hypothesis is H_t. (1933a, p.336)

In the decades that followed the publication of "the big paper" (as Neyman and Pearson referred to it) and the associated sequels (1933a, 1933b), the statistics world would grow to appreciate the enormous impact that this work would have on the concept of hypothesis testing and statistical inference at large. The mathematical justification for the use of the likelihood ratio, the designation of tests based on efficiency and power, and the development of acceptance procedures (specifically the christening of the two sources of error) have endured to the extent that it has led others to comment that "the mature theory of Neyman and Pearson is very nearly the received theory on testing statistical hypotheses" (Hacking, 1965, p.92).[2] What is important to appreciate, however, is that Neyman and Pearson did not consider themselves to be pioneers in the hypothesis testing landscape. They viewed their contribution as a supplement to Fisher's position, "a mathematical rounding off and improvement of Fisher's approaches" (Lenhard, 2006, p. 81). Naturally, Fisher interpreted this interference as unwarranted and responded in his signature polemic fashion: "A good deal of confusion has certainly been caused by the attempt to formalize the exposition of tests of significance in a logical framework different from that for which they were in fact first developed" (Fisher, 1971, p.26).

2.1.3 The Feud

The ungraciousness with which Fisher responded to this situation was ironic given that his present feud with the elder Pearson, Karl, had originated nearly two decades' previously when the upstart (undergraduate) Fisher took it upon himself to question, and venture to improve, Karl's method of moments theory in print. K. Pearson, editor of *Biometrika* – then the only statistics journal in existence – did not take kindly to the officiousness of this arriviste, and his counterattack launched an epic vendetta that persisted to his death (Spanos, 2013). Fisher was resolutely and systematically ostracized by the "statistical high priesthood" (Spanos, 2013):

[2] The author acknowledges that this is a dated reference; however, having been expressed three decades after the introduction of Neyman and Pearson's work, this certainly serves to justify the point being made about its legacy – arguably even if that quote no longer held in this era. Furthermore, even if one considers the modern NHST procedure to be more Fisherian in nature, there is no escaping its basic acceptance procedure structure (which of course was a Neyman-Pearson contribution).

invitation by the RSS was withheld and academic positions were either denied or offered with humiliating clauses[3]. It is unclear which of these injustices most fueled Fisher's resentment, but possibly in an attempt to punish E. Pearson for the sins of his father, Fisher took to openly criticizing the Neyman-Pearson school of thought.

To read Fisher is to draw the conclusion that there was no common ground between them. Firstly, he strenuously opposed the idea of a fixed, pre-determined level of significance. This is a point that he makes multiple times. First in his DOE textbook:

> Convenient as it is to note that a hypothesis is contradicted at some familiar level of significance such as 5% or 2% or 1% we do not, in Inductive Inference, ever need to lose sight of the exact strength which the evidence has in fact reached, or to ignore the fact that with further trial it might come to be stronger or weaker. The situation is entirely different in the field of Acceptance Procedures, in which irreversible action may have to be taken, and in which, whichever decision is arrived at, it is quite immaterial whether it is arrived at on strong evidence or weak. All that is needed is a Rule of Action which is to be taken automatically, and without thought devoted to the individual decision. (p. 25)

And later, in his *Statistical Methods and Scientific Inference* (1956): "No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" (p. 41).

Secondly, he took issue with the idea of *accepting* the null hypothesis. This is explained in his DOE textbook in 1935:

> The two classes of results which are distinguished by our test of significance are, on the one hand, those which show a significant discrepancy from a certain hypothesis;…and on the other hand, results which show no significant discrepancy from this hypothesis…In relation to any experiment we may speak of this hypothesis as the 'null hypothesis', and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. (DOE, p.16)

[3] Fisher was offered a professorship in Eugenics at University College, London in 1933 when K. Pearson retired but the contract included a clause that forbid him from teaching statistics. (Box, 1978, p.258)

In 1955, his argument had become vitriolic to the point that he even criticizes their use of the word *error* to describe such a situation (i.e., *accepting* the null hypothesis when it is *false*) arguing that label inadequate to describe the egregiousness implied therein: “it is a fallacy, so well known as to be a *standard example,* to conclude from a test of significance that the null hypothesis is thereby established; at most it may be said to be confirmed or strengthened” (Fisher, p. 73). To those who might argue, as Neyman and Pearson did, that rejection of a particular hypothesis is in fact evidence that the opposite must be true (thus giving credibility to the notion of acceptance), he has this to say in rebuttal:

> It might be argued that if an experiment can disprove [a hypothesis]…it must therefore be able to prove the opposite hypothesis…But this last hypothesis, however reasonable or true it may be, is ineligible as a null hypothesis to be tested by experiment, because it is inexact…it is easy to see that this hypothesis could be disproved by a single failure, but could never be proved by any finite amount of experimentation. It is evident that the null hypothesis must be exact…because it must supply the basis of the ‘problem of distribution’ of which the test of significance is the solution. (1971, p. 16)

Finally, and perhaps most resounding, was his vehement denial of error identification and specification, frequently referring to “errors of the second kind” as “logical fallacies” that are “committed only by those who misunderstand the nature and application of tests of significance” (Fisher, 1935). On this point, Fisher is unrelenting and repeats himself often. In 1935, he argues that the “notion of an error of the so-called ‘second kind’…has no meaning with respect to simple tests of significance, in which the only available expectations are those which flow from the null hypothesis being true” (p. 17). In 1955, he states that “the phrase, ‘errors of the second kind’, although apparently only a harmless piece of technical jargon, is useful as indicating the type of mental confusion in which it was coined” and goes on to add that “the fashion of speaking of a null hypothesis as ‘accepted when false’, whenever a test of significance gives us no strong reason for rejecting it…shows real ignorance of the research [worker]” (p. 73). In

1956, he softens the rhetoric just enough to allow an eloquent distillation of his fundamental disagreement to appear:

> The attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to hypothetical frequencies of possible statements, based on them, being right or wrong, thus seems to miss the essential nature of such tests. A man who ‘rejects’ a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases and when it is incorrect he will never be mistaken in rejection…however, the calculation is absurdly academic…The calculation is based solely on a hypothesis, which, in the light of the evidence, is often not believed to be true at all, so that the actual probability of erroneous decision, supposing such a phrase to have any meaning, may be much less than the frequency specifying the significance. To a practical man, also, who rejects a hypothesis, it is, of course, a matter of indifference with what probability he might be led to accept the hypothesis falsely, for in his case he is not accepting it. (p. 42)

In other words, the significance level provides an upper bound (rather than an approximation) on the Type I error probability because the null hypothesis is nearly *always* false. And in the case of the Type II ‘error’, the researcher has little use for the probability of a conclusion he has no intention of drawing. Furthermore, since the researcher never knows the truth of the hypothesis (in which case, the testing would be unnecessary), it is impossible to determine for any particular test whether an ‘error’ has been committed.

2.2 The Relationship Between P-values and Hypothesis Testing

2.2.1 Irreconcilable Differences

It would be easy to dismiss Fisher’s denigration of the Neyman-Pearson school as the polemical ranting of an irascible man, and in fact, there are those, Savage (1976), for example, who have done just that suggesting Fisher “never had sufficient respect for the work of that school to read it attentively” (p.448). But to do so, one disregards the churlish commentary Fisher endured (e.g. Neyman called some of Fisher’s work mathematically “worse than useless” (Nuzzo, p. 151)) and would miss the fundamental crux of the incompatibility.

Fisher (1955), contradictory evidence notwithstanding, makes the following claim:

> *I am casting no contempt* on acceptance procedures, and I am thankful, whenever I travel by air, that the high level of precision and reliability required can really be achieved by such means. But the logical differences between such an operation and the work of scientific discovery by…experimentation seem to me so wide that the analogy between them is not helpful. (p. 69, emphasis added)

However disinclined one might be to believe him on the previous point, what Neyman and Pearson believed to be a corrected and improved version of Fisher's work on tests of significance, he viewed as a "distortion" (Lenhard, 2006, p. 81) of his logic of inference – and on that point, he is on solid ground. The center of the disagreement was the distinction between *inductive inference* (Fisher) and *inductive behavior* (Neyman-Pearson); the semantical similarity between the two only serving to obscure the substantial divergence. Neyman held the point of view that a "logic of the uncertain such as is suggested by the phrase 'statistical inference' [was] illusory, but Fisher deplored that direction" and worked "fervently to establish a genuine theory of statistical inference" (Savage, 1976, p. 462).

In Fisher's estimation, there were two equally valid yet distinct "modes of human reasoning" (1932, p.257) in deductive and inductive logic. In the former, one argues "from the general to the particular"; whereas, in the latter, one argues "from the particular to the general" (i.e., from the sample to the population from which it was drawn). Fisher assumed, "as the experimenter always does assume, that it *is* possible to draw valid inferences from the results of experimentation" (1971, p. 3) and argued that "the mere fact that inductive inferences are uncertain cannot, therefore, be accepted as precluding perfectly rigorous and unequivocal inference" (1971, p. 4).

Fisher did not have a lot of support in this position. Mathematicians, trained almost exclusively in the technique of deductive reasoning, seemed unable or unwilling to make the distinction between uncertainty and lack of rigor, and mostly denied that rigorous inferences from the particular to the general were even possible (Fisher, 1935, p.39). The mathematicians' denial implied to Fisher their belief that "the process of learning by observation, or experiment, must always lack real cogency" (p.40) and he fought to dissuade them of this opinion. On the one hand, Fisher challenged them to remember the humble beginnings of deduction:

> It is as well to remember…that the principles and method of even *deductive* reasoning were probably unknown for several thousand years after the establishment of prosperous and cultured civilisations. We take a knowledge of these principles for granted only because geometry is universally taught in schools…Assuming the axioms, the body of their logical consequences is built up systematically and without ambiguity. Yet it is certainly something of an accident historically that this particular discipline should have become fashionable in the Greek Universities, and later embodied in the curricula of secondary education…since Euclid's time…unfettered individual judgment has been successfully denied in legal, moral, and historical questions, but…it has, none the less, survived, so far as purely deductive reasoning is concerned, within the shelter of apparently harmless mathematical studies. (1971, p.8-9)

On the other hand, Fisher explained that inductive reasoning is not new, just slow to be recognized. Arguing that men have always been capable of learning by experience, he claims "inductive inference is the only process known to us by which essentially new knowledge comes into the world" (1971, p.7). Experience might be imperfect and subject to immature reasoning, but that does not negate the potential for the "embryology of knowledge" (1971, p. 8). Once one has accepted that inductive inferences are legitimate in the context of experience, it is not so difficult to extend that consideration to experimentation. Letting Fisher explain: "Experimental observations are only experience carefully planned in advance, and designed to form a secure basis of new knowledge; that is, they are systematically related to the body of knowledge already acquired, and the results are deliberately observed, and put on record accurately" (1971, p. 8).

Neyman[4] had no use for Fisher's notion of inductive inference. Perhaps because he shared the sentiment of other mathematicians of his time, the whole N-P theory of hypothesis testing was based on the idea of establishing a method of behavior that was based solidly on the existing and widely agreed upon laws of probability, taking refuge in the secure mathematical haven of deductive logic. Neyman's response to Fisher was that some readers might think, as was the habit with Fisher's writings, "What an interesting problem is raised! How could I develop it further?" but that personally he was not amongst them and instead thought: "What an interesting way of asking and answering questions, but can't I do it differently?" (Neyman's discussion of Fisher's 1935 paper, p. 73). In a deft political move – sidestepping the question of whether legitimate inductive inferences could or should be performed without actually taking a stance[5] – Neyman essentially argued that the question was somewhat irrelevant given the ability to frame the whole of statistical estimation from a deductive perspective.

Fisher's motivation was to put forth significance testing and the use of his likelihood function as a key to the long-debated problem of induction. Neyman's motivation was to extend Fisher's ideas to include the use of the likelihood ratio in establishing acceptance procedures for inductive behavior. What is common to the approaches of both men is the notion of using experimental results in an evidentiary manner relative to one or more hypotheses. It is this common nucleus that led many to believe that reconciliation was possible, especially once the pair became colleagues at University College London; however, this thinking, as naïve now as it was then, ignores the underlying fundamental and philosophical incompatibility, personal feud

[4] This was originally a point of agreement for Neyman and Pearson; although it should be acknowledged that Pearson was not long for this philosophy and was quoted later as stating that "inductive behavior is Professor Neyman's field rather than mine" (1955).

[5] Incidentally, in later years, his views on the matter went from that of polite indifference/ cautious neutrality to that of blatant denial. See Neyman, 1957 for details.

notwithstanding[6]. According to Royall (1997), the differences in the approaches could be regarded as answering two different questions. In the case of Fisher, 'How should I interpret these observations as evidence?' and for Neyman and Pearson, 'What should I do, now that I have this observation?' (p. 64).

Fisher, developing his theories in the context of agricultural experimentation, believed "the purpose of significance testing was to provide the researcher with a precise solution to the problem of the generalizability of experimental outcomes, a problem he addressed in terms of inductive inference" (Halpin & Stam, 2006, p. 629). Tests of significance were used to make inference about the truth of the null hypothesis using an exclusive disjunctive position: *Either the null is false or an unlikely event has occurred, at most one of these is correct*. In Fisher's approach, the discrepancy between the single proposed hypothesis (i.e., the null) and the observed data is evaluated; one either rejects or fails to reject – but never accepts – the null on the sufficiency of the evidence.

In contrast, Neyman and Pearson did not frame their method explicitly in terms of experimentation, but rather in terms of providing a mathematical theory of statistical testing. Theirs was a decision theory instead of an inference theory, the purpose of which "may be regarded as the calculation of the power function of statistical tests and thus the selection of the best or most efficient test of a hypothesis" (Halpin & Stam, p. 632). In this approach, a dichotomous decision-making procedure was postulated based on two competing hypotheses such that a rule for inductive behavior was applied so that one hypothesis was rejected (and the

[6] Working together only further deteriorated the interpersonal relationship between the two. Described as "a bunch of squabbling children", Fisher and Neyman openly detested each other and took to publicly criticizing one another. Specifically, Neyman was said to "openly attack many of Fisher's ideas in lectures to his students" (Field, 2009, p.171).

alternative accepted by implication) whereby the results were not interpreted epistemically, but in terms of error frequencies in the long run.

It is the position of Lenhard (2006) that the key to these formative controversies between the two men was directly related to their conflicting views of mathematical modelling and its role in the statistical testing procedure. Fisher's "mediation" view stressed that "a model had to mediate between questions of application to real problems and data on the one hand, and questions of mathematical properties and arguments on the other" (p. 72). This was in contrast to the Neyman-Pearson "integration" view. Fisher believed that a statistician "uses mathematical reasoning within the logic of inference, e.g. building and adjusting a model to the data at hand and to the question under discussion" whereas in N-P theory, "the reasoning of the statistician himself, e.g. finding an appropriate acceptance procedure, has become an object of the mathematical argument" (p. 84).

For Fisher, modelling was the essence of the statistician's induction. Once proposed, a model paved the way for the use of deductive logic in that it "form[ed] the frame for formulating testable hypotheses" and "a plaT/Form for further mathematical argumentation" (Lenhard, 2006, p. 77); however, the critical component of this process was the inductive inference, i.e., the transition from concrete data to hypothetical model that enabled interpretation of that data.

> In short: the framing by means of a model is located at the beginning of the statistical treatment of a problem of application. 'The postulate of randomness thus resolves itself into the question, "Of what population is this a random sample?" which must frequently be asked by every practical statistician' (Fisher, 1922, p. 312-3).

Neyman and Pearson, in contrast, essentially "mathematized decision processes" (Lenhard, 2006, p. 85). Thinking about the problem as one of long-run consequences of repeated behavior as opposed to speculating about a single sample in hand, "the possible courses of action [could] be treated as mathematical objects that form[ed] risk sets with topological properties [i.e., convexity]" (Lenhard, 2006, p. 80). In this approach, the "model ha[d] to fit into the methodological framework that [was] conceived of…prior to modelling" (p.81). While Fisher believed that "modelling creates the objects one can argue about mathematically", Neyman-Pearson switched the role of inference logic by excluding hypothetical infinite populations and fixing the statistician's method through the use of reiterated procedures and competing hypotheses, essentially using mathematics to shape the situation in which modelling takes place (p. 84).

Neyman described modeling as a "theorizing step…taken in order to substitute for the phenomena an appropriate conceptual counterpart", the applicability of which was solely dependent upon the adequacy of the model (1955, p.16, 19). Then, as now, the assumption that phenomena could be modeled with "single valued functions of…specified arguments" was problematic and often generated models nowhere near comparable to the phenomena in question (p. 17). The solution was to use stochastic models, i.e., models in which "at least some of the intervening quantities [were] treated as random variables" (p. 17). In his 1955 article, Neyman clarifies his original position on *inductive behavior*, renaming it "the frequentists's theory of inductive inference" (p. 17). He claimed its suitable application is in those cases where "a stochastic model has been adopted to represent a given class of phenomena" (p. 17) and whose solution is "always unambiguously interpretable in terms of long range relative frequencies" (p. 19). Not wishing to squander an opportunity to reiterate his feelings for inductive behavior relative to inductive inference, he offers this statement on the relative suitability of the approaches:

> Specifically, if a problem of induction regarding phenomena *P*, for which there is a stochastic model *M*, is obtainable both within the frequentist and with a nonfrequentist theory of inductive inference, and if the latter is based on some elements *E* outside of the model *M* and , therefore, nonusable in the frequentist theory, then the correspondence of

> the nonfrequentist solution and the phenomena *P* is bound to be looser than the correspondence between the frequentist solution and the same phenomena. (p. 19)

Neyman, clearly frequentist himself, does not call out Fisher by name with this remark, but it is reasonably safe to assume that these comments could be interpreted as such. On a historical note, it is rather interesting that Neyman uses his frequentist philosophy as a means to distance himself from Fisher considering Fisher has been credited with pioneering modern frequentist statistics (Spanos, 2014). To be fair though, it is difficult to classify Fisher's philosophy given his propensity for criticism of both the Bayesian and frequentist approaches (e.g. Berger, 2003; Babu, 2012.). Ultimately, Fisher's criticism of frequentist testing was "its need for a fully specified alternative hypothesis, the associated difficulty of having to deal with a power function depending on unknown parameters", and that the errors in NP testing did not reflect the variation in evidence as the data ranged over the rejection regions (Babu, p. 4); Neyman criticized Fisher's p-values as violating the frequentist principle.

Fisher and Neyman, uncompromising in their respective positions, had not resolved their feud by the time of Fisher's death in 1962 and this conflict represented the "major antinomy in frequentist statistics" (Halpin & Stam, 2006, p. 626) in the 1960s. Some of the "most profound and ingenious efforts in [statistical theory] have gone into the search for new meanings for the concepts of inductive inference and inductive behavior" (Savage, 1961, p. 577); however, these efforts have been anything but successful. The adoption of statistical testing became an important but insurmountable challenge due to statisticians "simultaneously propounding two incommensurate approaches" (p. 626).

2.2.2 A Compromise is Made: Null Hypothesis Significance Testing

Researchers, growing weary of the feud, "lost patience and began to write statistics manuals for working scientists" (Nuzzo, 2015, p. 151). With Fisher now deceased, reconciliation with Neyman was impossible; thus “textbook authors were compelled to reconcile the foundational conflicts between the Fisher and Neyman-Pearson theories with the disciplinary goal of a single model of inductive inference” (Halpin & Stam, 2006, p. 627). This was a disastrous turn of events for the research community because these non-statisticians, lacking a thorough understanding of either approach, settled on the NHST procedure which would prove to be “a false rapprochement of the two theories” (Halpin & Stam, 2006, p. 627).

While quite understandable that such an amalgamation would be suggested owing to the fact that “both theories had enough similarities to be easily confused, especially by those less epistemologically inclined” (Perezgonzalez, 2015, p. 10), the reality is that the hybridization bastardizes both approaches and validates neither. In the words of Rozeboom (1997): “the use of *p*-values and null hypothesis testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students” (p. 335). Subject to condemnation almost immediately upon adoption, criticism ranged from the mild: “incoherent mishmash” (Gigerenzer, 1993) to the scathing: “a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring” (Meehl, 1967).

Despite its inappropriateness, NHST was readily adopted across disciplines and quickly became ubiquitous. As early as 1960, William Rozeboom observed that NHST had already attained the status of “religious conviction” (p. 416) despite the advancement of superior inferential techniques. His excoriation read: “despite the awesome pre-eminence this method has attained in our experimental journals and textbooks of applied statistics, it is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research” (p.416). Decades later and his words are as true as ever. Well-ingrained in the minds and practice of researchers, the controversial practice remains the standard operating procedure in psychology, the social sciences, and education (Perezgonzalez, 2015). "More interestingly", Perezgonzalez reports, is that "the numerous critiques raised against it for the past 80 years have not only failed to debunk NHST from the researcher's statistical toolbox, they have also failed to be widely known, to find their way into statistics manuals, to be edited out of journal submission requirements, and to be flagged by peer-reviewers" (p. 10).

What's wrong with NHST? "Well, among many other things, it does not tell us what we want to know and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (Cohen, 1994, p.997) Here Cohen is referring to the inverse probability fallacy, or "misapplication of deductive syllogistic reasoning" (p. 998), that occurs when researchers conflate P (H | D_0) with P (D | H_0). The *p*-value is a measure of the probability that the data (D) *could* have arisen *if* the null (H_0) were true; however, what the researcher really wants to know is the probability the null is true (H) *given* the data (D_0) – something altogether different entirely and incalculable with the present technique. As explained by Beaty and Torok (2014):

> The *p*-value is calculated under the assumption that the null hypothesis *is* true. It cannot therefore measure the probability of the null hypothesis. Any number of hypotheses may be true, including the null hypothesis. The *p*-value isn't the probability that **any** hypothesis is true. (emphasis in original, bolding added)

This misconception that the *p*-value provides a probability statement about the truth of either hypothesis is one of many misinterpretations and misunderstandings that stemmed from the unnatural hybridization of the two incompatible approaches to testing. Significant research (see for example Gissane, 2012; Cohen, 1994; Vidgen and Yasseri, 2016; Goodman, 2008; Ranstam, 1996) has been devoted to the consequences of this "uncomfortable marriage" (Goodman, 2008, p. 136), the common underlying factor of which is that "modern science has imposed the *p*-value on hypothesis testing, elevating its status and causing some confusion as to its meaning" (Gissane, 2012, p. 1).

Contrary to Fisher's original intention, the *p*-value was appropriated into the N-P acceptance procedure in a way that has caused it to be "dichotomized on an arbitrary basis…convert[ing] a probability into a certainty" (Vidgen & Yasseri, 2016, p.2). What was once an incremental accrual of evidence, a process in which the graduated appraisal of evidence had meaning, has been replaced with a decision-making procedure whose binary nature renders terms like *nearly* or *highly* significant nonsensical.

The problem with this dichotomization of hypothesis testing is that it has no "ontological basis" (Rosnow & Rosenthal, 1989, p. 1277). Furthermore, the test itself cannot decide the truth or falsity of any particular hypothesis. "While the scientist – i.e., the person – must indeed make decisions, his *science* is a systematized body of (probable) knowledge, not an accumulation of decisions" (Rozeboom, 1960). The problem here is the reduction of inference to a decision making exercise, which begs the question, "Does decision making fulfill the same purpose as statistical inference more broadly conceived?" (Gregoire, 2001, p. 4).

The answer to this largely rhetorical question can be found by looking no further than the sanctity attached to the 5 percent significance level. While the reasons for its selection are wholly logical – "both convenient and stringent enough to safeguard against accepting an insignificant result as significant" (Rosnow & Rosenthal, 1989, p. 1277) – the cutoff is itself completely arbitrary and, for the record, was *not* endorsed by Fisher or Neyman-Pearson, despite popular belief otherwise[7].

[7] In a contribution whose consequences could neither have been anticipated nor appreciated by Fisher himself, he included tables of distributions and particular quantiles in early editions of his SMRW textbook. This was not to glorify any particular cut-off but due in part to copyright disagreements with K. Pearson and because it was

Both camps propose purely hypothetical guidelines for significance (or none at all). The exclusion was quite deliberate although the reasoning behind this decision varied considerably between them. Fisher believed that "it [was] open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations [had] demonstrated a positive result" (1935, p. 13). For Neyman and Pearson, the decision – while still personal – was more complicated in that it required the researcher to choose the rejection region based on the simultaneous consideration of his chance of committing both types of errors. The balance must be adjusted between the risks P_I and P_{II} to meet the type of problem at hand and "determining just how the balance should be struck, must be left to the investigator" (1933a, p.296)

Cut to present day where "bright-line rules" dominate the landscape and the dictatorial role of *p*-values has all but eliminated genuine inference. With their famous expression, "surely God loves the .06 nearly as much as the .05" (1989, p. 1277), Rosnow and Rosenthal jest, but their argument for the interpretation of the strength of evidence for/against the null as a "fairly continuous function of the magnitude of p" is legitimate. While it should be fairly obvious that "conclusions do not become immediately 'true' on one side of the divide and 'false' on the other" (Wasserstein & Lazar, p. 9), mechanical rules demonstrate the practical irrelevance of the significance test in its failure to see genuine application to the inferential behavior of the researcher. As Rozeboom (1960) so eloquently put it, "what scientist in his right mind would ever feel there to be an appreciable difference between the interpretive significance of data, say, for which one-tailed $p = .04$ and that of data for which $p = .06$?" (p. 424). Unfortunately when p-

laborious to compile comprehensive tables by hand in that day. Pearson (1962) acknowledged that they also did not advocate a particular significance and in fact were "influenced by Fisher's tables of 5% and 1% because they lent themselves to the idea of choice, in advance of experiment, of the risk of the 'first kind of error' which the experimenter was prepared to take" (in Lehman, p. 1244).

values “become gatekeepers for whether work is publishable” (Wasserstein, 2016, p.1), this difference is not only appreciable but significant and represents just one of many misconceptions that researchers possess where *p*-values are concerned.

2.3 (Mis)Understanding *p*-Values

The *p*-value, informally, is “the probability under a specified statistical model that a statistical summary of the data…would be equal to or more extreme than its observed value” (Wasserstein & Lazar, 2016, p. 8). While statisticians and a good majority of scientists are able to recite this mantra in rote fashion, they are often “unwitting practitioners of an odd mix of Fisher & Neyman-Pearson ideas” (Gregoire, 2001, p.3). Research has shown that explaining it in intuitive terms or understanding the appropriate application of the concept remains elusive (Aschwanden, 2015) – even amongst scientists in the field and the majority of the misconceptions “flow from the unnatural union [of p-values and hypothesis testing]” (Goodman, 2008, p. 136).

2.3.1 Definition

Depending on the author and the audience, *p*-value definitions vary by textbook. Representing a blend of Neyman-Pearson and Fisherian interpretations – the balance of which is left up to the author – most definitions generally include three basic elements: an assumption of truth regarding the null hypothesis, an indication that the *p*-value is a probability, and some reference to the rarity of the observed sample statistic (Lane-Getaz, 2007, p. 6).

In Wasserman’s (2004) textbook for mathematicians, statisticians, and computer scientists, the *p*-value is presented by way of mathematical theorem (Theorem 10.12, p. 158):

> *Suppose that the size α test is of the form* reject H_0 if and only if $T(X^n) \geq c_\alpha$.
> *Then,* p-value = $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X^n) \geq T(x^n))$ *where x^n is the observed value of X^n.*
> *If $\Theta_0 = \{\theta_0\}$ then* p-value = $\mathbb{P}_\theta(T(X^n) \geq T(x^n))$.

Understanding that this definition, while precise, is difficult to interpret, he offers the following clarification: "We can express Theorem 10.12 as follows: The *p*-value is the probability (under H_0) of observing a value of the test statistic the same as or more extreme than what was actually observed" (p. 158). Perhaps as testament to the complexity in articulating the *p*-value, Wasserman does not stop with this theorem and its clarification, but in fact, offers multiple other interpretations:

- The *p*-value is the smallest level at which we can reject H_0. (p. 156)
- Informally, the *p*-value is a measure of the evidence against H_0: the smaller the *p*-value, the stronger the evidence against H_0. (p. 156)
- If H_0 is true, the *p*-value is like a random draw from a Unif(0,1) distribution. If H_1 is true, the distribution of the *p*-value will tend to concentrate closer to 0. (p. 158)

In Agresti & Franklin's (2009) textbook for general audiences, their presentation of the *p*-value lacks mathematical notation, but does include the three basic elements (p. 410):

> To interpret a test statistic value, we use a probability summary of the evidence against the null hypothesis, H_0. Here's how we get it: We presume that H_0 is true, since the "burden of proof" is on the alternative, H_a. Then we consider the sorts of values we'd expect to get for the test statistic, according to its sampling distribution. If the sample test statistic falls well out in a tail of the sampling distribution, it is far from what H_0 predicts. If H_0 were true, such a value would be unusual. When there are a large number of possible outcomes, any single one may be unlikely, so we summarize how far out in the tail the test statistic falls by the tail probability of that value and values even more extreme (meaning, even farther from what H_0 predicts)…This probability is called a **P-value**. The smaller the P-value, the stronger the evidence is against H_0.

As in the Wasserman text, multiple descriptions of the p-value are offered:

- The **P-value** is a tail probability beyond the observed test statistic value. Smaller P-values provide stronger evidence against the null hypothesis. (p. 410)
- The **P-value** is the probability that the test statistic equals the observed value or a value even more extreme. It is calculated by presuming that the null hypothesis H_0 is true. (p. 411)
- **P-value**: This is a probability summary of the evidence that the data provide about the null hypothesis. It equals the probability that the test statistic takes a value like the observed one or even more extreme.

- It is calculated by presuming that H_0 is true.
- The test statistic values that are "more extreme" depend on the alternative hypothesis. When H_a is two-sided, the P-value is a two-tail probability. When H_a is one-sided, the P-value is a one-tail probability.
- When the P-value is small, the observed data would be unusual if H_0 were true. The smaller the P-value, the stronger the evidence against H_0. (p. 459-460)

After looking at the way these texts present the *p*-value, it seems unsurprising that students and researchers would struggle with interpretation. Greenland, et al. (2016) describe this as a "key problem" that "there are no interpretations…that are once simple, intuitive, correct, and foolproof" (p. 337). The sheer magnitude of alternate presentations, offered for clarification, seem to obfuscate the understanding because it forces readers to make a choice between simultaneous reification of all interpretations or selection of the optimal representation from among a sea of distinct, but seemingly indistinguishable choices.

2.3.2 Misconceptions and Misunderstandings

In Goodman's 2008 article, he explains that defining the *p*-value is not equivalent to answering the questions *What does it mean?* and *What do people do with when they observe p < .05?* Researchers often fall victim to a whole host of misinterpretations and misunderstandings where the *p*-value is concerned because it is "elusive in meaning" (p. 136). Multiple attempts to catalog *p*-value misconceptions have been made by researchers from various disciplines. Three will be presented here, along with one set of principles for *correct* interpretation and application. These lists will be used to inform the test blueprint. That discussion, presented in Chapter 3 of this document, will exhibit a cross-tabulation of the misunderstandings, misconceptions, and misinterpretations presented thus far and will examine their suitability as it applies to the construct map.

2.3.2.1 Lane-Getaz's List

In 2007, in support of her instrument development project, Lane-Getaz compiled a list of correct conceptions and misconceptions of *p*-values and statistical significance as identified by her extensive literature review. This list cites many prominent researchers in the field and consolidates their individual efforts into an arguably exhaustive inventory on what was known (at that time) concerning *p*-value (mis)understanding. She further subdivided the items into four categories: Basic terminology and concepts; Relationships between inferential concepts; Logic of statistical inference; and Hypotheses, *p*-values, decisions, and error. See Table 2.1 (taken from Lane-Getaz, 2013, p. 22) for details.

Table 2.1 – Correct Conceptions and Misconceptions of *p*-values and Statistical Significance

Code	**Correct Conception (C) or Misconception (M)**
Basic terminology and concepts	
B-1	Demonstrating knowledge or confusion about basic language and concepts of inference (C or M)
B-2	Believing the p-value is always low (M)
Relationships between inferential concepts	
R-1	Confusing test statistics and p-values (M)
R-2	Confusing samples and populations (M)
R-3	Confusing α and Type I error rate or significance level with the p-value
R-4	Believing p-value is independent of sample size (M)
R-5	Believing reliability is 1 - p value (M)
R-6	Recognizing significance testing and confidence interval equivalence for means (C)
R-7	Confusing replications with sample size (M)
Logic of statistical inference	
L-1	Misusing the Boolean logic of contrapositive proof (M)
L-2	Misusing the Boolean logic of the converse (M)
L-3	Misinterpreting the p-value as the probability chance caused the observed results (M)
L-4	Misinterpreting the scope of inference; not attending to the study design (M)
L-5	Interpreting p-value as a conditional probability (C)
L-6	Checking necessary conditions for inference (C)
Hypotheses, p-values, decisions, and error	
H-1	Misinterpreting the p-values as the probability the alternative hypothesis is true (M)
H-2	Misinterpreting the p-values as the probability that accepting the alternative is false (M)
H-3	Misinterpreting the p-value as the probability the null hypothesis is true (M)
H-4	Misinterpreting the p-value as the probability the null hypothesis is false (M)
H-5	Interpreting p-values to make rejection decisions (C)
H-6	Confusing Type I and Type II error rates

2.3.2.2 Goodman's *Dirty Dozen*

In compiling her list, Lane-Getaz's motivation was to identify students' *p*-value misunderstandings in an effort to link them with learning outcomes for introductory statistics courses. In the year following the publication of Lane-Getaz's dissertation, Goodman released a similar list, but his had a different purpose and a different audience. In his 2008 article, he presented his list of twelve misconceptions in the name of service to the biomedical research community. Citing a survey of medical residents, in which only 62% were able to answer an elementary *p*-value interpretation question correctly (despite 88% of respondents self-reporting fair to complete confidence in interpreting *p*-values), this paper appears to be offered as a guide to what can (and does) get misinterpreted and misreported where *p*-values are concerned and offers a brief introduction to the use of Bayesian methods as a superior alternative to current inferential methods. He makes no claim to have exhausted the "range of misstatements about statistical measures, inference, or even the *p*-value" (p. 138) but claims that his list represents the "most common misinterpretations" (p. 136) and that the majority of those not listed are derivative from his list of twelve. See Table 2.2 for details.

Table 2.2 The Dirty Dozen (Goodman, 2008)

	12 P-value Misconceptions
1	If P = .05, the null hypothesis has only a 5% chance of being true.
2	A nonsignificant difference (e.g., P ≥ .05) means there is no difference between groups.
3	A statistically significant finding is clinically important.
4	Studies with P values on opposite sides of .05 are conflicting
5	Studies with the same P value provide the same evidence against the null hypothesis.
6	P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.
7	P = .05 and P ≤ .05 mean the same thing.
8	P values are properly written as inequalities (e.g., "P ≤ .02)" when P = .015)
9	P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.
10	With a P = .05 threshold for significance, the chance of a type I error will be 5%.
11	You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.
12	A scientific conclusion or treatment policy should be based on whether or not the P value is significant.

2.3.2.3 Guide to Misinterpretations (Greenland, et al.)

In 2016, in response to the ASA's *p*-value statement, Greenland and six of his colleagues compiled their own list, labeling it a "guide to misinterpretations". The objective was to adopt the format of Goodman's *Dirty Dozen* to generate "a list of misinterpretations" that can be used as a guide to "critically evaluate conclusions offered by research reports and reviews" (p. 340). With Goodman as a co-author, a handful of his original misconceptions made the cut, but the revised list was more extensive than the original, focusing on what *p*-values *don't* tell us, and expanded the topic coverage to include affiliated ideas such as confidence intervals and power. The authors claim that each item on their list has "contributed to statistical distortion of the scientific literature" and this guide is being presented as a way of "moving towards defensible interpretations and presentations" (p. 340) in that it be used as a "resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory may be limited but who wish to avoid and spot misinterpretations" (p. 337). Greenland and his colleagues sub-classified the items in this list into four categories: Common misinterpretations of single *p*-values, Common misinterpretations of *p*-value comparisons and predictions, Common misinterpretations of confidence intervals, and Common misinterpretations of power. See Table 2.3 for details.

Table 2.3 – Misinterpretations of *p*-values in research (Greenland, et al., 2016)

Category	Misinterpretation
Common misinterpretations of single P values	1. The P value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave P = 0.01, the null hypothesis has only a 1% chance of being true; if instead it gave P = .40, the null hypothesis has a 40% chance of being true. 2. The P value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the P value for the null hypothesis is 0.08, there is an 8% probability that chance alone produced the association. 3. A significant test result (P ≤ 0.05) means that the test hypothesis is false or should be rejected. 4. A nonsignificant test result (P > 0.05) means that the test hypothesis is true or should be accepted. 5. A large P value is evidence in favor of the test hypothesis. 6. A null hypothesis P value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated. 7. Statistical significance indicates a scientifically or substantively important relation has been detected. 8. Lack of statistical significance indicates that the effect size is small. 9. The P value is the chance of our data occurring if the test hypothesis is true; for example, P = 0.05 means that the observed association would occur only 5% of the time under the test hypothesis. 10. If you reject the test hypothesis because P ≤ 0.05, the chance you are in error (the chance your "significant finding" is a false positive) is 5% 11. P = 0.05 and P ≤ 0.05 mean the same thing. 12. P values are properly reported as inequalities (e.g., report "P < 0.02" when P = 0.015 or report "P > 0.05" when P = 0.06 or P = 0.70) 13. Statistical significance is a property of the phenomenon being studied, and thus statistical tests detect significance. 14. One should always use two-sided P values.
Common misinterpretations of P value comparisons and predictions	15. When the same hypothesis is tested in different studies and none or a minority of the tests are statistically significant (all P > 0.05), the overall evidence supports the hypothesis. 16. When the same hypothesis is tested in two different populations, and the resulting P values are on opposite sides of 0.05, the results are conflicting. 17. When the same hypothesis is tested in two different populations and the same P values are obtained, the results are in agreement. 18. If one observes a small P value, there is a good chance that the next study will produce a P value at least as small for the same hypothesis.
Common misinterpretations of confidence intervals	19. The specific 95% confidence interval presented by a study has a 95% chance of containing the true effect size. 20. An effect size outside the 95% confidence interval has been refuted (or excluded) by the data. 21. If two confidence intervals overlap, the difference between two estimates or studies is not significant. 22. An observed 95% confidence interval predicts that 95% of the estimates from future studies will fall inside the observed interval. 23. If one 95% confidence interval includes the null value and another excludes that value, the interval excluding the null is the more precise one.
Common misinterpretations of power	24. If you accept the null hypothesis because the null P value exceeds 0.05 and the power of your test is 90%, the chance you are in error (the chance that your finding is a false negative) is 10%. 25. If the null P value exceeds 0.05 and the power of this test is 90% at an alternative, the results support the null over the alternative.

2.3.3 ASA Statement on Statistical Significance and *p*-Values

The majority of research on the use of *p*-values is focused on pointing out the misconceptions, misunderstandings, and misinterpretations that scientists harbor. Statisticians, naturally concerned about the application of their methodological tools across disciplines, are aware of these *mis-es* and have been "sounding the alarm about these matters for decades, to little avail" (Wasserstein & Lazar, 2016, p. 130). The recently released *ASA Statement on Statistical Significance and P-Values*, breaks with the tradition of identifying what is wrong with the way people conduct statistical inference and elects instead to "establish firm general principles which focus on what is right rather than what is wrong" (Spiegelhalter, 2017, p. 41). Controversy among the participants drafting the statement included concern about its lasting impact, especially considering its "striking vagueness", but there was support for a policy statement penned in positive terms, as expressed by Goodman: "We need to formulate a vision of what *success* looks like…" (Matthews, 2017, p. 40).

After many iterations, the final statement, intended to "provide the community a service" (Wasserstein & Lazar, 2016, p. 129), ignores, among other things, alternative hypotheses, error types, and power (p. 130) – although even that decision was a point of contention among those involved. In the introduction, the ASA is precisely clear about what the statement is – "a few select principles that could improve the conduct or interpretation of quantitative science" – and what it is not – "this statement does not seek to resolve all the issues relating to sound statistical practice, nor to settle foundational controversies" (p. 131). The statement's six principles are the following:

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

The ASA statement is arguably not prescriptive enough; however, its authors are transparent and unapologetic about that point. The authors, in belief that "the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value" (Wasserstein & Lazar, 2016, p. 131), declared that the statement merely "articulates in nontechnical terms a few select principles" (p. 131), selected according to consensus in the statistical community, with the aim of drawing "renewed and vigorous attention to changing the practice of science with regards to the use of statistical inference" (p. 130).

2.3.4 Gender Differences in Statistical Reasoning

With all of the attention surrounding the *p*-value controversy, and all of the research dedicated to explaining how and why *p*-value interpretation is so elusive, it might be tempting to think of this as a universal affliction. This notion, however, is not entirely accurate. The extant literature would suggest that understanding *p*-values is difficult for everyone, but harder for some than others. Gender studies and cross-cultural studies have indicated that there are statistically significant differences in subgroup comparisons on measures of both correct statistical reasoning and misconceptions – with these effects being, at best, only partially attributable to differences in coursework or cognitive ability. For the purposes of this dissertation, the discussion will be limited to studies investigating gender differences.

Liu (1998, as cited in Garfield, 2003) used the Statistical Reasoning Assessment (SRA, more details about this instrument to follow in Section 2.4.1) to compare the correct reasoning and misconception scores of males and females among college students in the USA and Taiwan. In both samples, the female subgroup had significantly higher *misconception* scores than their male counterparts; furthermore, the male subgroup outscored the female subgroup on *correct reasoning* in both samples (although not a statistically significant difference in the USA).

Similarly, Tempelaar, Gijselaers, and Schim van der Loeff (2006) administered the SRA in a cross-cultural study comparing the performance of Dutch students with that of international students (i.e., other European countries), specifically investigating gender differences. The results were not only consistent with that of Liu, but provided evidence that "the gender effect is rather substantial [*p*-values < 0.005]: males score more than 5% higher in total correct reasoning, and more than 9% lower in total misconceptions, than females, with Cohen's *d* effect size ranging between small and medium" (Section 3.1, para. 4).

In 2017, Martin, Hughes, and Fugelsang examined the joint effects of gender and experience on statistical reasoning, again using the SRA as the primary performance measure. Results are commensurate with Tempelaar, Gijselaers, and Schim van der Loeff: "As predicted, males performed better than females overall…scoring higher on the CR (i.e. *correct* reasoning) scale by an average of 12 percentage points across experience levels… *p* < .001 …and scoring lower on the MISC (i.e. *misconceptions*) scale by an average of 7 percentage points across experience levels… *p* < .001" (p. 463). As expected, experience was directly related with performance, but they did not find this improvement moderated the gender gap. For instance, "only women having taken at least two courses in statistics reached the level of performance of men with *no* experience in statistics" (p. 463, emphasis added).

It is important to bear in mind that in none of these studies were *p*-value misconceptions investigated *per se*, but rather statistical reasoning at large, and therefore the generalizability of these results to the present study is peripheral at best. That being said, what is compelling and relevant to this dissertation is the evidence suggesting that females, across countries and levels of experience and programs of study, not only underperform relative to males in general, but particularly acutely where *mis*conceptions are concerned.

2.4 Research Instruments to Assess *p*-Value Understanding

The literature is replete with articles and research studies expounding on the difficulties that exist where *p*-value interpretation is concerned – a subset of which have been previously cited in this dissertation; however, despite evidence that these misunderstandings are both common and persistent, there is a remarkable dearth of reliable and/or valid instruments for their measurement. As observed by Lane-Getaz (2007), “there is no single instrument used across these studies to measure the various difficulties cited in the literature” (p. 20; see also, Sotos, et al., 2007). As such, the research proposed here aims to fill that void by producing an instrument for that purpose.

2.4.1 Existing Instruments

The majority of existing instruments measuring statistical literacy and reasoning are essentially comprehensive – that is, they are designed to capture student understanding of a wide range of topics typically covered in service courses for non-statisticians. The most notable of these is the Statistical Reasoning Assessment (SRA) developed by Garfield and Konold in 1998 as part of the National Science Foundation (NSF)-funded ChancePlus Project. It is included in this discussion because of its status as the first instrument designed to assess students’ ability to understand statistical concepts and apply statistical reasoning (Garfield & Chance, 2000). A multiple-choice test consisting of 20 items, the assessment was again unique in that it measured both misconceptions and correct understandings. Table 2.4 (Garfield & Chance, 2000) summarizes the content coverage of the SRA. The SRA was initially used on high school students, but administration has expanded to include college students and has been adapted for use in other countries. A pioneering endeavor, the SRA has been cited and employed in countless research studies; however, despite its popularity and ability to "provide some useful information regarding the thinking and reasoning of students as they solve statistical problems", it is – in the words of its author – "problematic as a research and evaluation tool" owing to disappointing reliability and validity diagnostics.

Table 2.4. Topic Coverage of the SRA (Garfield & Chance, 2000)

Correct Reasoning Skills and Misconceptions Measured by the SRA	
Correct Reasoning Skills	1. Correctly interprets probabilities
	2. Understands how to select an appropriate average
	3. Correctly computes probability
	a. Understands probabilities as ratios
	b. Uses combinatorial reasoning
	4. Understands independence
	5. Understands sampling variability
	6. Distinguishes between correlation and causality
	7. Correctly interprets two-way tables
	8. Understands importance of large samples
Misconceptions	1. Misconceptions involving averages
	a. Average is the most common number
	b. Fails to take outliers into consideration when computing the mean
	c. Compares groups based on their averages
	d. Confuses mean with median
	2. Outcome orientation misconception
	3. Good samples have to represent a high percentage of the population
	4. Law of small numbers
	5. Representativeness misconceptions
	6. Correlation implies causation
	7. Equiprobability bias
	8. Groups can only be compared if they are the same size

In a follow-up endeavor, Garfield and Gal launched the ARTIST project (Assessment Resource Tools for Improving Statistical Thinking), also funded by the (NSF), with the purpose of addressing the need "to develop reliable, valid, practical, and accessible assessment instruments" (delMas, Ooms, Garfield, & Chance, 2006, p. 1). The resources created and compiled by the research team serve to assist faculty teaching statistics across disciplines with tools to evaluate students and improve their courses. Through the website, instructors can access three achievement measures (the Comprehensive Assessment of Outcomes in Statistics, the Statistics Concept Inventory, and the aforementioned SRA), the ARTIST Topic Scales, and an extensive item database where they can create customized assessments.

The ARTIST Topic Scales are a set of eleven online unit tests, ranging from seven to twelve items each, designed to cover "an intersection of topics included in most introductory statistics courses" (delMas, et al., 2006, p. 2). The CAOS instrument is a 40-item test to measure statistical literacy and reasoning. Topic coverage includes "basic literacy and reasoning about descriptive statistics, probability, bivariate data, and basic types of statistical inference" with the "intent to develop a set of items that students completing any introductory statistics course would be expected to understand" (delMas, et al., 2006, p.2). The SCI (Statistics Concept Inventory), developed by Kirk Allen, was designed to identify student misconceptions and assess student understanding of topics typically encountered in an introductory statistics course in a manner consistent with how the FCI (Force Concept Inventory) measured physics students' knowledge of Newton's Law. Items were designed based on the AP statistics curriculum and were selected from textbooks and statistics journals; factor analysis has suggested content coverage in four categories: Descriptive Statistics, Probability, Inferential Statistics, and Graphical.

Each of the aforementioned assessments was designed to be used with undergraduate populations. The Statistical Literacy Assessment Scale (SLAS) is different than these other instruments in that the target audience was graduate students (initially, and later to include college statistics teachers). The SLAS, developed by Enriqueta Reston (2005) "as a classroom tool for assessing statistical literacy of graduate students who have diverse undergraduate backgrounds, but who have taken at least a 3-unit introductory statistics course in their undergraduate coursework" (p. 1), is comprised of 15 items along two dimensions: "(1) understanding of basic statistical concepts and terminology used in the context of real-world situations and (2) understanding of claims and arguments based on data from various media outlets" (p. 1). The assessment is of mixed-response format requiring the students to make a closed-ended selection (Yes/No/Cannot Decide) to each prompt and then justify or explain their answer.

As these instruments illustrate, the majority of existing instruments approach the topics of statistical literacy and reasoning broadly. The subject of inference, including hypothesis testing and *p*-values, is reliably present in these assessments, but not the sole topic under investigation. Notable exceptions include the ARTIST Test of Significance topic scale and the RPASS (Reasoning about *p*-values and Statistical Significance) instrument. Unlike the ARTIST topic scale which was designed to be a component of a larger set, the RPASS was designed as a stand-alone measure of *p*-value understanding.

The RPASS was developed by Lane-Getaz (2007) in direct response to a documented deficit in the way of research instruments dedicated to the assessment of understanding and misunderstanding of *p*-values and statistical significance. This instrument, currently in its 11th version, consists of 38 items that are a combination of T/F, multiple-choice, and open-response format. Item content was selected to reflect correct conceptions and misconceptions of *p*-values and statistical significance as culled from the research literature and is sub-divided into four categories (as presented in Table 1 of this document).

2.4.2 Inadequacies of Existing Instruments to Address the Present Need

As enumerated in the previous section, there have been some attempts at assessment instruments to measure statistical literacy and/or reasoning; however, it is necessary to articulate why these measures are inadequate to address the research questions under investigation in this study. The aforementioned existing instruments suffer from one or more of the following limitations where the present study is concerned: inadequate psychometric properties, alternative population, lack of singular focus, and/or insufficient topic coverage.

The SRA was designed for, and tested on, high school students and undergraduates in an introductory course. The topic coverage was intentionally broad and therefore not focused on statistical inference in general or *p*-values in particular. Due to the lackluster psychometric properties of the instrument as a whole, there is no reason to believe that the subset of items measuring *p*-value understanding would perform sufficiently well as a stand-alone measure.

The ARTIST resources are similarly inappropriate for use in the current study. The CAOS, like the SRA, was a comprehensive measure aimed at undergraduates in a first course in statistics. Despite good overall reliability in a national sample (Cronbach's α = .82), there is no evidence to suggest that the 14 inference-related items can be used as an isolated subscale (Lane-Getaz, 2007, p. 39). Looking at the Test of Significance topic scale, this measure does meet the singular focus criterion; however, reliability and validity evidence has not been released. The SCI, was a comprehensive measure aimed at specifically undergraduate engineering students, and the generalizability of its use outside of the target population and the parent institution have been called into question. Furthermore, the suggestion of an underlying factor structure supporting a Statistical Inference subscale has since been undermined upon subsequent analysis (Allen, 2006).

The SLAS is unique to this list of research instruments in that it was designed specifically for a graduate student population. That being said, data was collected exclusively with students in a course for *teachers*, and thus results are not necessarily generalizable to graduate students from other disciplines. The items on this assessment are related to the proposed context of this research in that the respondents are asked to evaluate the research findings and claims made by others; however, there was no specific mention of *p*-values or hypothesis testing and the content leaned more towards sampling issues and graphical interpretations.

The RPASS instrument is the most closely aligned with the aims of the present research. This assessment was designed to solely measure misunderstanding where *p*-values and statistical significance are concerned; thus, having a singular focus allows the psychometric properties to be interpreted specifically for this measure of inference (as opposed to examining a subset of inference items among a larger instrument). The most current reliability and validity evidence exhibit no signs of concern where psychometric properties are concerned as all reported indices are quite satisfactory. Potential limitations to use of this instrument are few, but there are two worth noting. Firstly, although some graduate students were included in early rounds of pilot/field testing of the RPASS, the measure was primarily intended and tested with undergraduate students. Secondly, and perhaps most importantly, the topic coverage of the instrument might be insufficient to address the current need. While the test blueprint was based on an extensive literature review and was designed to capture all correct conceptions and incorrect misunderstandings identified heretofore, the RPASS has been used primarily in pre/post studies to evaluate the influence of particular curricular innovations in ameliorating misconceptions (see Lane-Getaz, 2017) and therefore the content is closely aligned with learning outcomes for introductory courses – which may or may not align with the requisite knowledge for submitting and reviewing journal articles.

2.5 Summary and Research Questions

This literature review has provided evidence that *p*-value misconceptions are widespread and persistent owing to foundational controversies, a convoluted relationship with significance testing, and an increase in usage by non-statisticians. The rampant misuse has been attributed to the replicability crisis and other related consequences. In response to the banning of *p*-value reporting by *The Journal of Basic and Applied Psychology*, the ASA released their policy statement with principles for prudent usage of *p*-values in an attempt at a course correction for the research community where statistical inference is concerned. It was the intention of the ASA to draw "renewed and vigorous attention to changing the practice of science" (Wasserstein & Lazar, 2016, p. 130). Two years on and it remains to be seen if this tactic will have the desired effect, but it invites investigation as to whether or not researchers understand and can(are) apply(ing) the statement's principles in practice.

While research on *p*-value misinterpretation abounds, the focus is typically on explaining how the misconceptions arose, enumerating the misunderstandings, and describing the consequences of misuse. What is noticeably absent from these studies however is the existence of a valid and reliable instrument to measure said misinterpretations, Lane-Getaz's RPASS being a notable exception. This literature review indicates that existing assessments suffer from one or more of the following limitations where the present study is concerned: inadequate psychometric properties, alternative population, lack of singular focus, and/or insufficient topic coverage. This suggests the need to build an instrument to assess the extent to which doctoral students, i.e., future researchers, harbor misinterpretations that could statistically distort scientific findings and are incompatible with the ASA guidelines. To that end, this dissertation investigates two research questions: firstly, *Can a sufficiently reliable and valid measure of p-value misinterpretations be constructed*?, and secondly, *What do the results of the KPVMI-1 administration tell us about the current level of p-value fluency among doctoral students nationally?*

Chapter 3 – **Methods**

The primary purpose of this research study is to determine the extent to which doctoral students, i.e., future researchers, struggle with interpreting and reporting *p*-values in the context of independent research and peer review. As identified in Chapter 2, misconceptions in this area abound, yet many of these taxonomies do not necessarily reflect the population of interest in this particular study. Furthermore, while much research has been conducted that *identifies p*-value misconceptions, considerably less progress has been made in terms of instrument development for their assessment. To date, there exists no widely-accepted and psychometrically validated instrument to measure this 'construct', partially because the construct is not well-defined nor necessarily unidimensional. Independent efforts have produced instruments to measure related constructs (e.g. ARTIST topic scales) and some of these contain items related to *p*-values (e.g., RPASS); however, an instrument aimed at the target population (future researchers) and written in the appropriate context (independent research / peer review) does not exist. To that end, in order to satisfy the main research objective, the present study is necessarily two-pronged: first, an assessment instrument was created and validated, and second, upon verification of satisfactory resolution of the first research objective, the instrument was then utilized to make claims about the performance of the population under investigation.

In this chapter, the details for both prongs of the research design will be described and justified. First, the research setting and nature of participants will be discussed. Following that is a description of the instrument creation and validation plan. The final section of this chapter outlines the plan for instrument administration and data analysis.

3.1 Overview of the Study

3.1.1 Research Setting and Nature of Participants

Research in the area of statistical literacy, reasoning, and thinking has been primarily limited to that of undergraduate populations and introductory statistics courses. The current research wishes to expand the extant literature by targeting graduate students, and specifically, PhD candidates. It is important to study this population, i.e., *future* researchers, because the *Basic and Applied Social Psychology* journal's ban on *p*-values called into question the methodological competence (or lack thereof) of *practicing* researchers – precisely the group responsible for the education and evaluation of the future researchers' statistical methodology.

The exact nature of the participants and the requisite number evolved across all three phases of this research (instrument creation, validation, and administration). This information is described in appropriate detail in each of the corresponding sections of this chapter, but in general, the population under investigation consisted of doctoral students enrolled at colleges and universities in the United States (not in the field of Statistics) who had completed a level of methodological coursework appropriate to their discipline.

Throughout the study, responses obtained from those students who had completed all methodological coursework were prioritized; however, participation was not restricted to include *only* those students. Since not all students complete their program in a strictly linear fashion (with all coursework ending before research and exams begin), it was decided that it could be unfairly prejudicial to exclude students merely on the basis of outstanding courses on their plan of study – courses that may or may not even be methodological. In the instrument creation and development phases of the study, recruitment efforts targeted students near the completion of their programs (and therefore closer to becoming practicing researchers). In the later phases, specifically the field test and the national sample, any willing participant who met the prerequisite training requirement (i.e., active enrollment in a PhD program) was considered in the interest of assuaging sample size concerns.

3.1.2 Phases of the Study

The first research question, *Can a sufficiently reliable and valid measure of p-value misinterpretations (in a research context) be constructed*?, was addressed via the development and validation of the *Keller P-value Misinterpretation Inventory* instrument (KPVMI-1). This endeavor occurred across three phases.

During Phase I, a test blueprint was generated for the instrument based on identified misinterpretations from the literature. An initial item pool, PV0, was generated and subjected to expert review by members of the dissertation committee. Modifications were made in accordance with advice from the statistical advisors and the preliminary version of the instrument, PV1, was created.

During Phase II, the instrument was piloted with Virginia Tech (VT) students (n = 15). Results from cognitive labs (a form of think-aloud interview) and a usability study (n = 5 and n = 10, respectively) informed a revised version of the instrument, PV2.

During Phase III, the instrument was twice field tested: first using students enrolled in a large interdisciplinary graduate research methods course here at Virginia Tech (n = 79), and later in a national sample (n = 207). Instrument validation (to include item analysis and reliability/validity diagnostics) occurred at this stage. A subset of items was identified to serve as the (current) final version of the instrument (PV3).

The second research question, *What do the results of the KPVMI-1 administration tell us about the current level of p-value fluency among doctoral students nationally?*, was addressed via analysis of a subset of the field test data (n = 147) with respect to performance on the subset of items considered sufficiently validated as developed in Phases I-III (*KPVMI*-1). Inferences about respondents' *p*-value fluency were drawn from this data, in totality and by subgroup where appropriate.

3.2 Instrument Creation and Validation

3.2.1 The 4 Building Blocks

The methodological plan for the instrument development was guided by the "4 Building Blocks" approach as outlined by Wilson (2005). In Wilson's view, "an instrument is always something secondary: there is always a purpose for which an instrument is needed and a context in which it is going to be used" (p. 6). The coordination of the purpose, context, and content is achieved by means of an iterative cycle that begins with a construct map, proceeds to items and item scores, and concludes with measures. The cycle is permitted to repeat as often as necessary for refinement.

The first of the building blocks, the *construct map*, is the tool that provides precision and clarity to the *construct*, defined as "the theoretical object of our interest in the respondent" (Wilson, 2005, p. 6). The most important features of the construct map are that there is a "coherent and substantive definition for the content of the construct" and "an idea that the construct is composed of an underlying continuum" (p. 26). Under the assumption that the construct can be modeled by a unidimensional, continuous variable, the construct map delineates the relative position of the respondents based on their level of ability and the responses based on the sophistication of performance exhibited. The construct map helps the researcher to differentiate between respondents who have "more" or "less" of the construct under consideration and to determine the gradient between "novice" and "expert".

The second of the building blocks, items design, is the manifestation of the theoretical construct in real-world situations. As Wilson explains, "the construct is not clearly defined until a large set of items has been developed and tried out with respondents" (2005, p. 42). The items must be designed in an intentional way in order to prompt informative responses that reveal the respondent's tendency or level of the particular characteristic under investigation. An instrument is the "result of a series of decisions that the measurer has made regarding how to represent the construct or, equivalently, how to stratify the 'space' of items…and then sample from those strata" (p. 45).

While an undeniable relationship intuitively exists between the first two building blocks, "the construct and the items are both only vaguely known…and causality is often unclear at this point, perhaps the construct 'causes' the responses that are made to the items" (Wilson, 2005, p. 12). In prematurely abbreviated instrument development endeavors, the conceptual approach ends here and this results in several shortcomings: "arbitrariness in choice of items and item formats, no clear way to relate empirical results to instrument improvement, and an inability to use empirical findings to improve the idea of the construct" (p. 12). To avoid these consequences, the third and fourth building blocks embody an inferential process in which responses to the items are used to infer the underlying construct.

The third building block, the outcome space, is the first step in the inference and involves "mak[ing] a decision about which aspects of the response to the item will be used as the basis for the inference, and how those aspects of the response are categorized and then scored" (Wilson, 2005, p. 13). The goal of this step is to discover and understand the qualitatively different ways in which students respond to a task and bring "order and sense" to this likely large and "probably unruly bunch of potential responses" (p. 67). The product of the outcome space is a scoring guide or rubric.

The fourth building block, the measurement model, is the second step in the inference and involves the translation of scored responses to locations on the construct map. As Wilson explains, "the measurement model must help us to understand and evaluate the scores that come from the item responses and hence tell us about the construct, and it must also guide the use of the results in practical applications" (2005, p. 16). The measurement model may be formal or informal in nature and does not necessarily have to be statistical; however, it must satisfy two requirements (p. 116):

1. The measurement model must enable one to interpret the distance between respondent and response on the construct map, and
2. The measurement model must enable one to interpret distance between different responses on the construct map and also the difference between different respondents

Collectively, the four building blocks provide both a path for inference about a construct and a guide to instrument construction as well. Using this approach, the construct is first defined as embodied in the construct map, tasks are developed that engage the construct, responses to items are categorized and scored, and finally the measurement model is applied to analyze the scored responses (Wilson, 2005, p. 18). As the "measures can be used to reflect back on the success with which one has measured the construct" (p. 18), this process is necessarily cyclic and often repetitive. In the next sections, the specific application of the four building blocks to the present study will be discussed as they unfold throughout the phases.

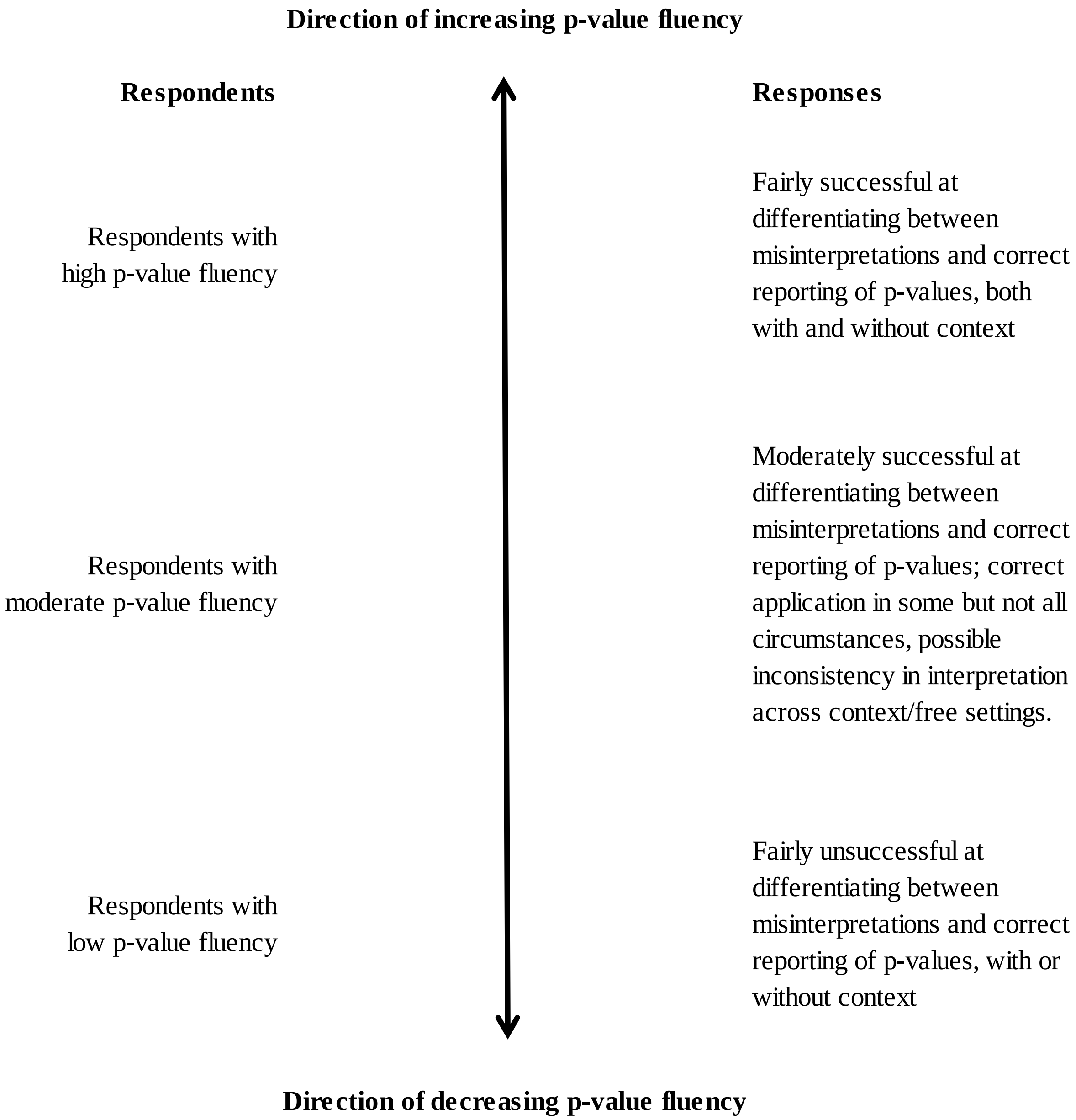
3.2.2 Phase I

Phase I of the study addressed the first two of these building blocks: the construct map and items design.

3.2.2.1 The Construct Map

In the present study, the overarching construct is the ability to interpret and report *p*-values in a manner consistent with the ASA's principles; however, this construct is clearly multidimensional and is therefore not a good candidate for a construct map approach in its current form. Accordingly, this research adopts *p-value fluency* as the construct under investigation and, while leaving open the possibility of mapping a correspondence between sub-scale scores and the six ASA principles in a subsequent analysis, currently assumes the position (until evidence to the contrary is produced) that this construct is unidimensional. *P-value fluency*, as defined by this research, will be the ability to use *p*-values in a manner that is methodologically defensible and does not "contribute to the statistical distortion of the scientific literature" (as defined by Greenland, et al., 2016). This is discussed in greater detail in the item-writing and test blueprint sections of this document.

The current instrument is a performance, rather than deficits, model – the key distinction being that in the latter, items are designed to capture misconceptions and in the former, items are designed to measure appropriate usage and understanding. Accordingly, the construct map (see Figure 3.1), diagramming both responses and respondents, represents the continuum of *p-value fluency* such that higher scores correspond with more fluency (as opposed to in the deficits model where higher scores would correspond to more misconceptions).



Direction of increasing p-value fluency

Respondents

Responses

Respondents with high p-value fluency

Fairly successful at differentiating between misinterpretations and correct reporting of p-values, both with and without context

Respondents with moderate p-value fluency

Moderately successful at differentiating between misinterpretations and correct reporting of p-values; correct application in some but not all circumstances, possible inconsistency in interpretation across context/free settings.

Respondents with low p-value fluency

Fairly unsuccessful at differentiating between misinterpretations and correct reporting of p-values, with or without context

Direction of decreasing p-value fluency

Figure 3.1. Construct Map of p-value Fluency

3.2.2.2 The Internal and External Models

Theory-based models are an essential component of the instrument development process. The theoretical framework on which the instrument is based includes the construct map and its affiliated internal and external models. The internal model describes the relationships between the elements of the construct, decomposing the construct into its various components. In doing so, the internal model moves the initial construct map toward test specifications by making explicit the internal structure of the instrument. The external model situates the construct in context, in other words, describes its relationship with affiliated constructs. Additionally, these models lay the foundation for various aspects of validity arguments.

Existing models of statistical literacy and reasoning define the construct in terms of the expectation for an average citizen or the goals of an introductory course. These models, like the curricula modeled after them, aim to educate consumers, not producers, of statistics. For researchers who are – and future researchers (PhD students) who will be – producers of statistics, general statistical literacy is obviously necessary, but certainly insufficient. Research scientists must be held to a standard of statistical proficiency that is commensurate with their influence.

In short, quantitative research proficiency is comprised of those aspects of a statistical disposition most pertinent to a scientist's career: communicating statistical results to others and contradicting claims made without proper statistical foundation. At the most basic level, a scholar must possess research skills in two main areas, here described by Schuyten (1991):

> 1. Skills needed to read and *evaluate* surveys, experiments, and other studies dealing with substantive problems in the research area.
> 2. Skills needed to do research while planning a study, analyzing the data, *interpreting and generalizing the results*. To this second category '*reporting the results'* can be added. Both categories of skills rely on statistical and methodological competencies. (emphasis added)

In contrast with the layperson who must be just fluent enough to digest the condensed, simplified statistics made public, the researcher must possess the ability to navigate the interplay between his conceptual framework and the methodological and statistical competence that will allow him to analyze data that was collected in the real world but must ultimately be discussed in terms of his theoretical world.

For most researchers, they become acquainted with this process while earning their PhD. It is during this training that they are exposed to the substantive theory of their chosen field, while also accumulating methodological and statistical knowledge. Here at Virginia Tech, this is how the graduate school describes this undertaking:

> "The Doctor of Philosophy program is designed to prepare a student to become a scholar: that is, to discover, integrate, and apply knowledge, as well as to communicate and disseminate it… A well-prepared doctoral student will have the ability to understand and critically evaluate the literature of the field and to apply appropriate principles and procedures to the recognition, evaluation, interpretation and understanding of issues and problems at the frontiers of knowledge" (Council of Graduate Schools, 2005, p.1).

The process of earning a PhD involves the acquisition of very specialized content knowledge; however, that is not the sole purpose, or arguably even the true purpose, of the education. Based on the previous definition, the true objective of a doctoral program is that of a research apprenticeship.

Writing one's dissertation allows the student researcher to demonstrate competency in doing independent research; a process that involves the activation and application of the methodological and statistical knowledge accrued thus far in training. However, it often requires that which goes beyond the coursework to knowledge that is imposed by the research questions and thus the student researcher begins the transition to researcher as he/she learns to update his/her statistical knowledge as applicable to the situation at hand. Practicing researchers must be fluent with all phases of statistical inquiry: design, data collection, data analysis, and reporting of results – both for their own work and in evaluating the work of others. Furthermore, they must be capable and willing to update and adapt as the methodological techniques of the field change in order that they are not only utilizing best practice in their own work but able to effectively peer review.

The external model proposed in this research situates *p*-value fluency as a sub-component of *Research Proficiency* – one of two primary components of the PhD Plan of Study. See Figure 3.2 for details.

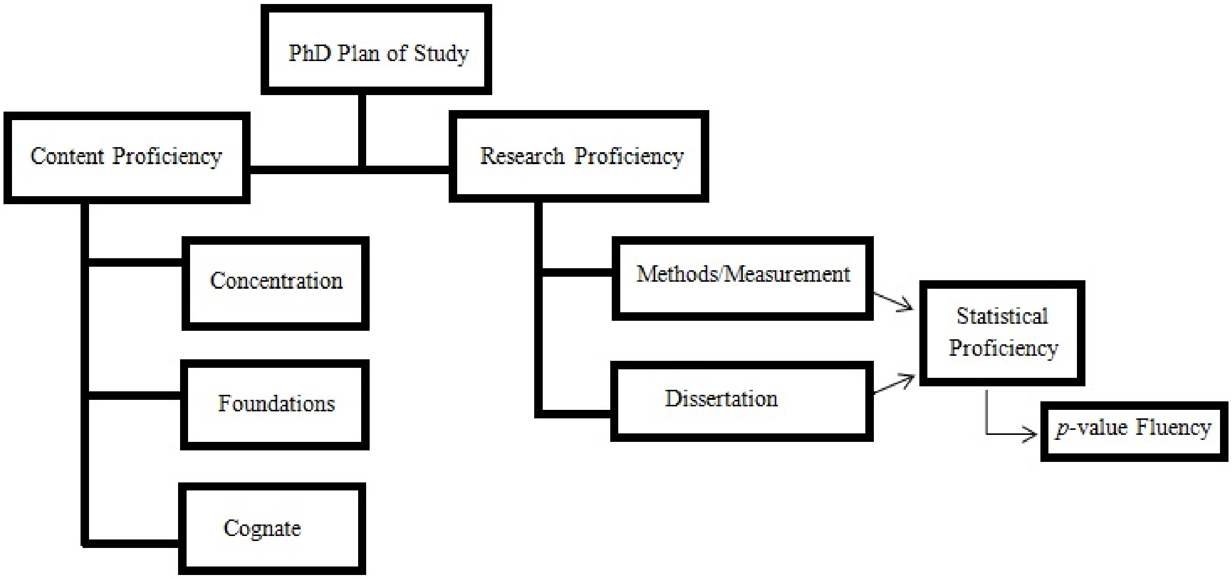


Figure 3.2. External Model Situating *p*-value Fluency in the PhD Plan of Study

The statistical knowledge a researcher must possess is that which is necessary for performing and evaluating research in a way that is complementary to the research profession. *P-value fluency*, is a subset of this statistical knowledge, and as defined by this research, is the ability to use *p*-values in a manner that is methodologically defensible and does not "contribute to the statistical distortion of the scientific literature" (as defined by Greenland, et al., 2016). As presented in Chapter 2, a significant amount of research has been dedicated to cataloguing and classifying misconceptions and misinterpretations where p-values are concerned. These lists tend to overlap and comprehensiveness varies by author – both of these factors contribute to the difficulty in producing a single evaluative assessment to measure this (in)ability. For the sake of this research, three of the more prominent lists have been cross-referenced for compatibility and redundancy. Table 3.1 presents these misinterpretations alongside the misconceptions and difficulties of Goodman and Lane-Getaz, respectively, organized by sub-topic.

Given that the population of interest in this study is future researchers, the list generated by Greenland and his colleagues was selected as the basis of the topic coverage (i.e., the internal model) as that seemed the most appropriate choice considering that list was based on errors committed by practicing researchers. Furthermore, their list was the most exhaustive and did subsume most of the items presented in the other studies.

Table 3.1 – Reorganization of p-value Misunderstandings, Misconceptions, and Misinterpretations

	Misinterpretation	Misconception	Difficulty
Basic Ideas	11. P = 0.05 and P ≤ 0.05 mean the same thing. 12. P values are properly reported as inequalities (e.g., report "P < 0.02" when P = 0.015 or report "P > 0.05" when P = 0.06 or P = 0.70) 13. Statistical significance is a property of the phenomenon being studied, and thus statistical tests detect significance.	7. P = .05 and P ≤ .05 mean the same thing. 8. P values are properly written as inequalities (e.g., "P ≤ .02" when P = .015)	B-1: Confusion about basic language and concepts of inference
Test statistics & p-values	9. The P value is the chance of our data occurring if the test hypothesis is true; for example, P = 0.05 means that the observed association would occur only 5% of the time under the test hypothesis. 14. One should always use two-sided values.	6. P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis. 11. You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.	L-5*: Interpreting the p-value as a conditional probability R1: Confusion between test statistics and P-values B-2: Believing the p-value is always low.
Statistical Significance and Effect Size	7. Statistical significance indicates a scientifically or substantively important relation has been detected. 8. Lack of statistical significance indicates that the effect size is small. 6. A null hypothesis P value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.	3. A statistically significant finding is clinically important. 12. A scientific conclusion or treatment policy should be based on whether or not the P value is significant. 2. A nonsignificant difference (e.g., P ≥ .05) means there is no difference between groups.	

	Misinterpretation	Misconception	Difficulty
P-values & Chance	2. The P value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the P value for the null hypothesis is 0.08, there is an 8% probability that chance alone produced the association.		L-3: Misinterpreting the p-value as the probability chance caused the observed results; probability due to chance
Type I & Type II Errors	10. If you reject the test hypothesis because P ≤ 0.05, the chance you are in error (the chance your "significant finding" is a false positive) is 5%	9. P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%. 10. With a P = .05 threshold for significance, the chance of a type I error will be 5%.	R-3: Confusion between significance level, α, and Type I error rate and the P-value H-6: Confusing Type I and Type II error rates
P-values and Power	24. If you accept the null hypothesis because the null P value exceeds 0.05 and the power of your test is 90%, the chance you are in error (the chance that your finding is a false negative) is 10%. 25. If the null P value exceeds 0.05 and the power of this test is 90% at an alternative, the results support the null over the alternative.		L-4: Misinterpreting the scope of inference; not attending to the study design R-7: Confusion of replications with sample size
P-values across Studies	15. When the same hypothesis is tested in different studies and none or a minority of the tests are statistically significant (all P > 0.05), the overall evidence supports the hypothesis. 16. When the same hypothesis is tested in two different populations, and the resulting P values are on opposite sides of 0.05, the results are conflicting. 17. When the same hypothesis is tested in two different populations and the same P values are obtained, the results are in agreement. 18. If one observes a small P value, there is a good chance that the next study will produce a P value at least as small for the same hypothesis.	4. Studies with P values on opposite sides of .05 are conflicting. 5. Studies with the same P value provide the same evidence against the null hypothesis.	R-2: Confusion between samples and populations R-4: Believing P-value is independent of sample size

	Misinterpretation	Misconception	Difficulty
P-values and the Truth or Falsity of Hypotheses	1. The P value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave P = 0.01, the null hypothesis has only a 1% chance of being true; if instead it gave P = .40, the null hypothesis has a 40% chance of being true. 3. A significant test result (P ≤ 0.05) means that the test hypothesis is false or should be rejected. 4. A nonsignificant test result (P > 0.05) means that the test hypothesis is true or should be accepted. 5. A large P value is evidence in favor of the test hypothesis.	1. If p = .05, the null hypothesis has only a 5% chance of being true.	H-3: Misinterpreting P-value as the probability the null hypothesis (H0) is true H-1: Misinterpreting P-value as the probability the alternative hypothesis (Ha) is true H-2: Misinterpreting P-value as the probability that accepting the alternative hypothesis (Ha) is false H-4: Misinterpreting P-value as the probability the null hypothesis (H0) is false H-5*: Interpreting p-values to make rejection decisions L-1: Misusing the Boolean logic of inverse (a→b confused with ~b→ ~a) to interpret a hypothesis test (confusion of the inverse) L-2: Misusing the Boolean logic of converse (a→b replaced with b→a) (confusion of the converse)
P-values and Confidence Intervals	19. The specific 95% confidence interval presented by a study has a 95% chance of containing the true effect size. 20. An effect size outside the 95% confidence interval has been refuted (or excluded) by the data. 21. If two confidence intervals overlap, the difference between two estimates or studies is not significant. 22. An observed 95% confidence interval predicts that 95% of the estimates from future studies will fall inside the observed interval. 23. If one 95% confidence interval includes the null value and another excludes that value, the interval excluding the null is the more precise one.		R-6*: Recognizing significance testing and confidence interval equivalence for means

In keeping with Schuyten's presentation of a scholar's research skills has belonging to two categories, namely conducting and evaluating research, the internal model proposed by this research similarly sub-divides the construct of *p*-value fluency. A researcher must understand the appropriate implications and interpretations of the *p*-value in a generic, context-free setting in order to apply them appropriately in their work. Additionally, the researcher must be able to evaluate the statements made by fellow researchers and determine if the conclusions, provided in the context of the study's setting, flow logically and defensibly from the statistical results obtained. This ability to interpret and report *p*-values in both contextual and context-free settings extends to all of the sub-categories identified by Greenland and his colleagues (as seen in Table 1) and thus the structure of the internal model (Figure 3.3) reflects this.

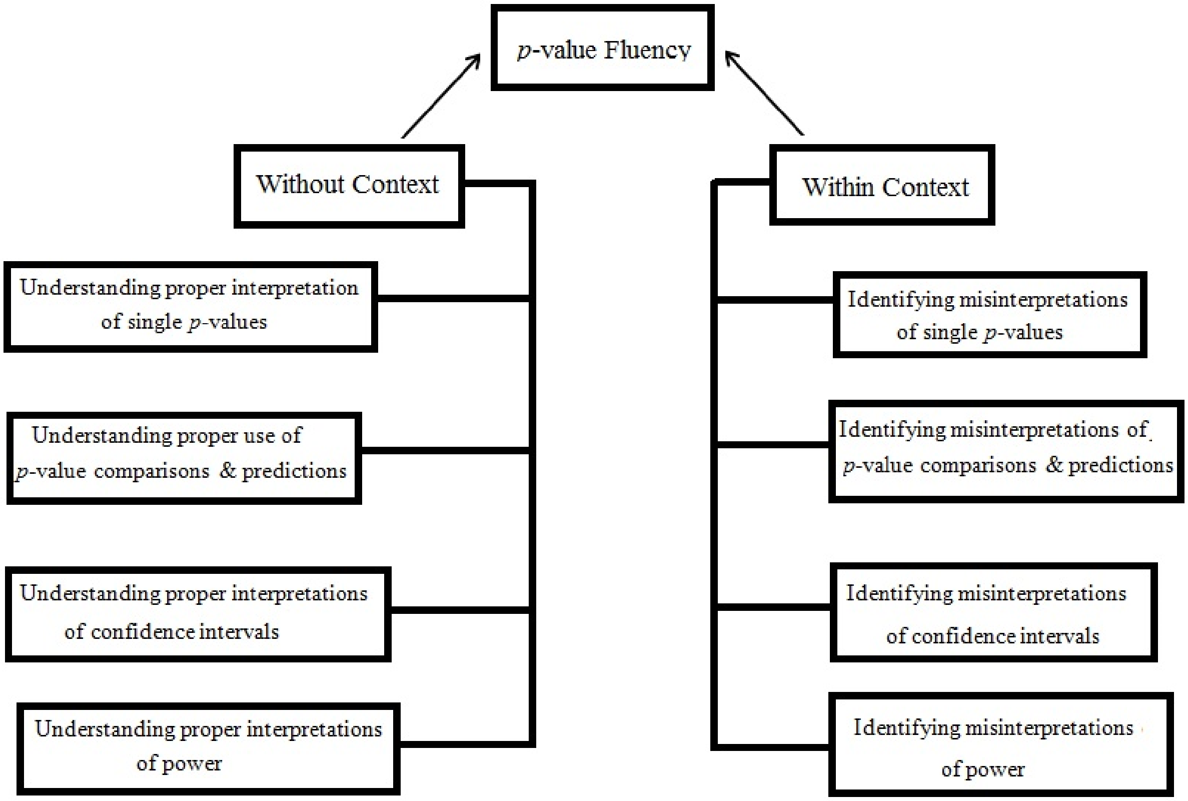

p-value Fluency
Without Context
Within Context
Understanding proper interpretation of single *p*-values
Understanding proper use of *p*-value comparisons & predictions
Understanding proper interpretations of confidence intervals
Understanding proper interpretations of power
Identifying misinterpretations of single *p*-values
Identifying misinterpretations of *p*-value comparisons & predictions
Identifying misinterpretations of confidence intervals
Identifying misinterpretations of power

Figure 3.3. Internal Model of p-value Fluency

3.2.2.3 The Items Design

In instrument development, the construct map and the internal model guide the item-writing phase of the instrument development process that begins with an initial item pool and culminates with the test blueprint. In this research, Greenland et al.'s list of misinterpretations (Table 1) was adopted as the original internal model upon which the initial item pool (PV0) was to be generated. Upon closer examination, amidst researcher concerns regarding the compatibility of the internal model to the proposed construct (i.e., *p-value fluency)* and skepticism as to dimensionality, the internal model was updated to include not the list of 25 misinterpretations in its entirety, but a subset of 18. The categories labeled *Misinterpretations of Confidence Intervals* and *Misinterpretations of Power* were removed on the basis of these being somewhat peripheral concepts to that of p-values themselves. See Figure 3.4 for details.

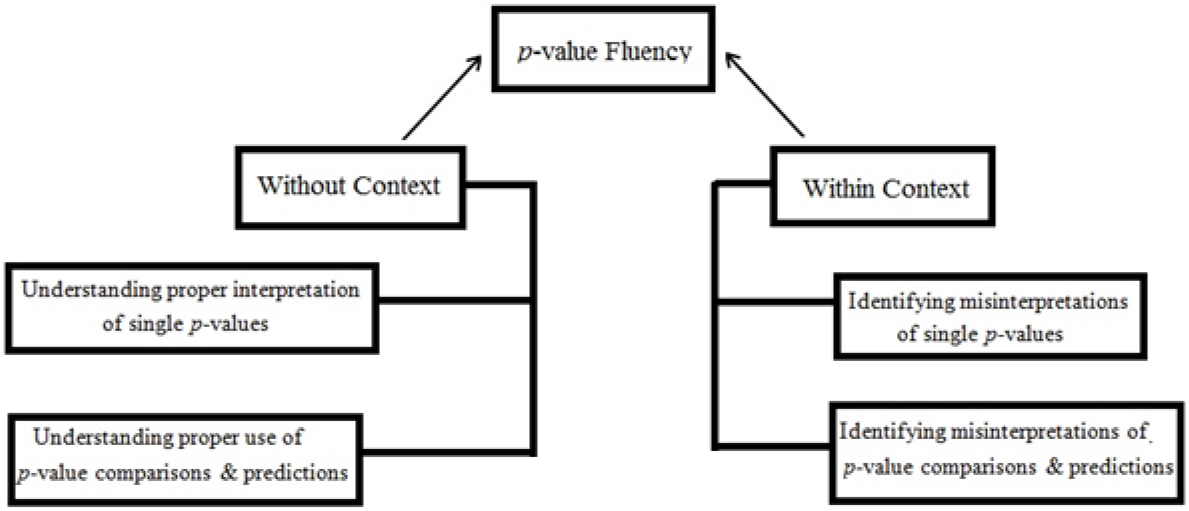

p-value Fluency
Without Context
Within Context
Understanding proper interpretation of single *p*-values
Understanding proper use of *p*-value comparisons & predictions
Identifying misinterpretations of single *p*-values
Identifying misinterpretations of *p*-value comparisons & predictions

Figure 3.4. Internal Model for PV Instrument Item Development, Revised

For each misinterpretation, the original goal was that a minimum of four items was to be generated – two based on the original statement (context-free) and at least two based on research vignettes (in context). The context-free items would be used to determine which misinterpretations are held by the members of the target population *themselves;* whereas, the research vignette items would be used to assess the ability of this population to identify the *misinterpretations of others* in the context of a peer review.

Item quality and suitability was determined by use of an item panel. Based on the guidelines presented in the Appendix to Chapter 3 in Wilson (2005, p. 59-61), an item panel was convened consisting of content and measurement experts for the purpose of a thorough professional review of the items before testing them with actual respondents from the target population. The panelists, both VT faculty members serving on the dissertation committee, were selected based on their expertise and proximity to both the subject matter and the target population. One of the panelists is a faculty member in the department of Educational Research and Evaluation specializing in instrument design; the other is a faculty member in the department of Statistics. Both teach quantitative research methods courses for graduate students pursuing degrees in fields other than Statistics. The panelists were given advance copies of the construct map and the items along with an overview of the framework for the context of the planned instrument. At the panel meeting, the items were reviewed, one-by-one as necessary, in order to answer questions, address comments, and note suggested revisions raised by the panel members. Specific details by item were solicited, but also a general sense for the comprehensiveness of the instrument relative to the construct was discussed. Specifics of the item panel review process as well as individual item revisions and iterations are presented in the Results section.

The input of the item panel was used to create and modify the test blueprint. This organizational tool outlines the scope and specificity of the instrument. The test blueprint is organized by content type and process level and dictates the number of items from each combination of content/process level that the final version of the instrument will require. In practice, many more items need initially be generated to account for the attrition that will arise during validation. After revising current items based on the panel's recommendations, supplemental items were created as necessary in order that the item pool contained approximately twice as many items as the blueprint specifications. The iterative nature of this process dictated that repeat panels be convened when necessary to review supplemental items and evaluate structural modifications to the test blueprint.

It is important to note that at this stage of the instrument development process, the outcome space had not been determined and therefore the process levels in the proposed test blueprint were both preliminary and speculative. The evolution of the test blueprint in response to the pruning of the item pool and panel feedback is discussed in detail in the Results section.

Table 3.2 – *Proposed Test Blueprint for the PV Instrument*

	Process Level			
Item Content	**Context-Free**		**In Context**	
	Affirmative	Negative	Correct	Incorrect
Misinterpretations of single P-values	14	14	14	14
MI1	1	1	1	1
MI2	1	1	1	1
MI3	1	1	1	1
MI4	1	1	1	1
MI5	1	1	1	1
MI6	1	1	1	1
MI7	1	1	1	1
MI8	1	1	1	1
MI9	1	1	1	1
MI10	1	1	1	1
MI11	1	1	1	1
MI12	1	1	1	1
MI13	1	1	1	1
MI14	1	1	1	1
Misinterpretations of P-value comparisons and predictions	4	4	4	4
MI15	1	1	1	1
MI16	1	1	1	1
MI17	1	1	1	1
MI18	1	1	1	1

3.2.3 Phase II – Pilot Study

Phase II of the study addressed the third building block: the outcome space. "Empirical evidence", as Wilson explains (2005, p. 69), is an essential part of the pilot investigation of an instrument and can be used to "support the ordering of an outcome space". For the present study, this evidence was gathered by way of a small pilot test that was supplemented with cognitive interviews and a usability study.

Before administering the instrument on a large scale, a pilot study was conducted as a means of providing an additional opportunity to revise the items and implementation procedures as necessary in order to obtain the most accurate and usable data. In addition to providing a sense of how the instrument will function in practice, another “particularly useful” aspect of pilot studies is their ability to make “quantitative estimates of response rates…and help in setting sample sizes for the full study” (Dillman, et al., 2009, p. 228). Pilot studies achieve these aims by targeting the same population as intended for the ultimate data collection but on a smaller scale.

The current pilot study was conducted in two phases, both of which utilized purposeful samples of VT students. The respondents, selected on a volunteer basis, were asked to participate in either a cognitive lab (Phase IIB, target sample size n = 5) or the usability study (Phase IIA, target sample size n = 10).

3.2.3.1 Phase IIA

In PhaseIIA, the instrument was administered online using Qualtrics survey software. This usability study portion of the pilot study was utilized to estimate response time and determine challenges individuals faced when registering responses. This included, but was not limited to details such as font size, number of items per page, arrangement of distractors, presence of progress bar, ease of accessibility to survey link, etc. This information was assessed via an exit interview questionnaire which was included as an addendum to the end of the instrument.

The usability study participants were recruited based on satisfactory completion of a capstone course in quantitative methods (EDRE 6106 and STAT 5616, in particular). The reason for this criterion was that these students would have been exposed to an appropriate minimum level of methodological training, and enrollment in these courses is limited to graduate students outside the field of statistics. Email solicitation yielded 13 participants. Participation was accepted on a first-come, first-served basis and was terminated when the target sample size had been exceeded due to financial considerations. Participants were financially compensated for their time and participation on a sliding scale (usability study participants were compensated $5 each for completion of Parts I and II of the instrument, and $5 for completion of the exit questionnaire). This phase of the research was completed under the supervision of the Virginia Tech Institutional Review Board (VT IRB) and all pertinent documentation therein (i.e., approval letter, recruitment materials) can be found in Appendix A.

3.2.3.2 Phase IIB

In PhaseIIB, cognitive labs, also referred to as "think-aloud" interviews were used to gather additional empirical evidence in support of the outcome space. In this type of investigation, a small, but representative, sub-sample of the target population (n = 5 for the present study) is asked to complete the current version of the instrument while actively describing their thought process as they respond "as a means of determining whether respondents comprehend questions as intended by the survey sponsor and whether questions can be answered accurately" (Dillman, et al., 2009, p. 221). In this semi-structured environment, the researcher can actively inquire of the participant or simply record and notate observations. In either case, the data is used both to determine if items are functioning as desired and also to delineate the range of sophistication of potential responses.

The cognitive lab participants completed the same version of the assessment as the usability study participants (the online Qualtrics version), but were video and audio recorded while doing so. The participants were encouraged to "think-aloud" as they answered the items. This commentary provided insight into the participant's level of statistical reasoning while also supplying feedback regarding aspects of the user experience. The former was used to determine if the instrument was able to distinguish between respondents at varying ability levels, while the latter was used to improve the usability of the instrument itself. The recorded sessions were transcribed for later analysis, selected details of which can be seen in the Results section.

Cognitive lab participants were personally recruited based on selection criteria involving progress towards dissertation (i.e., at least 3rd-year), program of study (i.e., non-Statistics majors required to use quantitative methods in their research), suitability to the nature of the data collection process (i.e., ability and willingness to extensively articulate their thought processes), and familiarity with the research team. All five targeted individuals agreed to participate. Cognitive lab participants were compensated at a rate of $15/hour. This phase of the research was completed under the supervision of the Virginia Tech Institutional Review Board (VT IRB) and all pertinent documentation therein (i.e., approval letter, recruitment materials) can be found in Appendix B.

Upon completion of the pilot study, a revised version of the instrument, PV2, was constructed. Modifications were made in accordance with participants' usability concerns and cognitive interview responses, as well as information gathered pertaining to response and completion rates. The evolution of the test blueprint as it pertains to these modifications is discussed further in the Results section..

3.2.4 Phase III – Field Test

Phase III of the study addressed the fourth building block: the measurement model. A field test was conducted in order to compile reliability and validity evidence in pursuit of the instrument validation plan devised in the tradition of the Messick (1989, 1995) 6-component framework (details to follow in 3.2.5). The field test portion of this research was conducted in stages which included both an internal sample of VT students and an external sample of doctoral students across the United States. Data from both samples were used individually and collectively as appropriate to the validation efforts.

3.2.4.1 Phase IIIA

Phase IIIA was the local field test. The class roster of the Spring 2018 (VT) section of STAT5616 was the basis for the sampling frame. Justification for this recruitment decision included the following considerations: (1) the large class size suggested a reasonable response rate would yield an adequate number of participants, (2) the course is the capstone of a year-long methods sequence and thus the students should possess the requisite knowledge to engage meaningfully with the instrument, and (3) the course is a service course offered to graduate students across departments and would theoretically yield a diverse representation of programs of study.   The study was advertised in class and also by email. Incentivization was offered in the form of course extra credit. The instrument was administered online via Qualtrics survey software. The target sample size was n = 100 and recruitment efforts yielded 79 completed responses spanning 37 programs of study. This phase of the research was completed under the supervision of the Virginia Tech Institutional Review Board (VT IRB) and all pertinent documentation therein (i.e., approval letter, recruitment materials) can be found in Appendix C.

3.2.4.2 Phase IIIB

Phase IIIB was the national field test, undertaken in the interest of robustness and generalizability of the findings outside of VT. Since the target population was doctoral students pursuing programs of study outside Statistics, the decision was made to recruit participants at PhD-granting institutions only; and more specifically, at R1 institutions (as classified by the Carnegie Classification of Institutions of Higher Education), because it is reasonable to assume that the highest proportion of future researchers would be trained at the institutions generating the most research. This listing resulted in a population of 115 institutions shown here in Table 3.3.

Recruitment at these institutions was handled indirectly. Soliciting support from graduate student organizations (such as student government associations), members of the executive boards were contacted in the Fall semester of 2018 via email and asked to disseminate the study advertisement to graduate students at their institutions. No guaranteed financial compensation was promised to participants; however, in accordance with social exchange theory as outlined in Dillman, Smyth, and Christian (2009), a multi-stage prize raffle was employed as inducement. Within each participating institution, one prize winner was selected to receive a \$10 electronic gift card to the merchant of their choice (Starbucks, Amazon, or iTunes). Across all participants, eight prize winners were selected to receive \$50 electronic gift cards to the merchant of their choice (again, Starbucks, Amazon, or iTunes). The instrument was administered online via Qualtrics survey software. There were no minimum thresholds for an institution's participation; any and all valid[8] responses, from any R1 institution, were considered. The target sample size was n = 400 in order to generate stable parameter estimates. Recruitment efforts yielded 207 responses spanning at least 16 unique institutions[9]. This phase of the research was completed under the supervision of the Virginia Tech Institutional Review Board (VT IRB) and all pertinent documentation therein (i.e., approval letter, recruitment materials) can be found in Appendix D.

[8] Valid responses would not include non-doctoral students or students pursuing a degree in Statistics.

[9] Only 87 respondents identified their institutional affiliation. For the remaining 120 respondents, there was no way to recover this information and thus no way to state with accuracy exactly how many institutions actually participated.

Table 3.3 – Sampling Frame of R1 Institutions

R1 Institutions

Arizona State University-Tempe
Boston College
Boston University
Brandeis University
Brown University
California Institute of Technology
Carnegie Mellon University
Case Western Reserve University
Clemson University
Colorado State University-Fort Collins
Columbia University
Cornell University
CUNY Graduate School and University Center
Duke University
Emory University
Florida International University
Florida State University
George Mason University
George Washington University
Georgetown University
Georgia Institute of Technology (main campus)
Georgia State University
Harvard University
Indiana University-Bloomington
Iowa State University
Johns Hopkins University
Kansas State University
Louisiana State University
Massachusetts Institute of Technology
Michigan State University
New York University
North Carolina State University - Raleigh
Northeastern University
Northwestern University
Ohio State University (main campus)
Oregon State University
Pennsylvania State University (main campus)
Princeton University
Purdue University (main campus)
Rice University
Rutgers University-New Brunswick
Stanford University
Stony Brook University
SUNY at Albany
Syracuse University
Temple University
Texas A & M University-College Station
Texas Tech University
The University of Tennessee-Knoxville
The University of Texas - Arlington
The University of Texas - Austin
The University of Texas - Dallas
Tufts University
Tulane University of Louisiana
University at Buffalo
University of Alabama at Birmingham
University of Arizona
University of Arkansas
University of California-Berkeley
University of California-Davis
University of California-Irvine
University of California-Los Angeles
University of California-Riverside
University of California-San Diego
University of California-Santa Barbara
University of California-Santa Cruz
University of Central Florida
University of Chicago
University of Cincinnati-Main Campus
University of Colorado Boulder
University of Connecticut
University of Delaware
University of Florida
University of Georgia
University of Hawaii at Manoa
University of Houston
University of Illinois at Chicago
University of Illinois at Urbana -Champaign
University of Iowa
University of Kansas
University of Kentucky
University of Louisville
University of Maryland-College Park
University of Massachusetts-Amherst
University of Miami
University of Michigan-Ann Arbor
University of Minnesota-Twin Cities
University of Mississippi
University of Missouri-Columbia
University of Nebraska-Lincoln
University of New Mexico-Main Campus
University of North Carolina at Chapel Hill
University of North Texas
University of Notre Dame
University of Oklahoma-Norman Campus
University of Oregon
University of Pennsylvania
University of Pittsburgh-Pittsburgh Campus
University of Rochester
University of South Carolina-Columbia
University of South Florida (main campus)
University of Southern California
University of Utah
University of Virginia (main campus)
University of Washington-Seattle Campus
University of Wisconsin-Madison
University of Wisconsin-Milwaukee
Vanderbilt University
Virginia Commonwealth University
Washington State University
Washington University in St Louis
Wayne State University
West Virginia University
Yale University

The original intention of this research was to use the local field test (VT sample) for instrument validation and to use the national sample in investigating the second research question. Smaller than expected sample sizes rendered this approach difficult, and at times impossible; thus, the decision was made to postpone certain aspects of the validation plan until both sets of data were in hand and to pool them when appropriate and/or necessary.

Due to the iterative nature of the process, instrument validation is never ‘complete’ but continually undertaken with each administration of the assessment. As such, certain components of the validation plan (e.g. external validity) were not undertaken until after Phase IIIB, while other diagnostics (e.g. classical item analysis) were performed on Phase IIIA data and then repeated after Phase IIIB. A further discussion of this appears in the Results section.

3.2.5 Instrument Validation

A validation plan for this instrument was devised in the tradition of the Messick (1989, 1995) 6-component framework. In his words, “Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores” (1992, pg. 1487). To that end, evidence will be collected in each of the six categories: Content, Substantive, Structural, Generalizability, External, and Consequential; to piece together a complete picture of the construct being measured.

3.2.5.1 Content Validity

This category refers to evidence of content relevance, representativeness, and technical quality. The proposed instrument has a clearly defined purpose: to measure *p*-value fluency in PhD candidates, and has a defensible rationale for doing so: as a screening mechanism for future academicians to determine if they possess the requisite skill set to participate as a fully-functioning member of the research community. In keeping with the notion of research preparedness, the test specifications address both facets of the research process: conducting research and evaluating research across an extensive, research-based list of practitioner misinterpretations.

When conducting research, the researcher must understand the valid implications of the methodology employed and limitations therein and be able to select from a myriad of possible conclusions to report the most appropriate for the data and the situation; furthermore, when reviewing research, the reviewer must argue for or against the current methodology or interpretations of others. It is for these reasons that the item types were developed as they were, i.e., context-free and with context (i.e., in the evaluation of researcher actions), in order to simulate authentic conditions.

The instrument was developed using the construct modeling framework for instrument development as presented in Wilson (2005). The scope of the assessment was determined by a panel of experts in *p*-value research and further, the individual items were repeatedly subjected to review by academicians responsible for teaching research methods to graduate students. These subject-matter experts are familiar with both the types of knowledge necessary for this type of work and also the frequently misunderstood topics therein.

In keeping with classical test theory, the items were examined for technical quality by use of the following indicators: item difficulty and item discrimination. Item difficulty, as measured by the proportion of respondents who answered a given question correctly, should ideally be close to .5; particularly low or high values provide the least amount of discrimination. Item discrimination for criterion-referenced assessments should be positive; negative values indicate that the question is likely disproportionately deceptive or misleading to those respondents who are scoring at the high level of the construct.

All analysis was run using either jMetrik (Meyer, 2018) or IRTpro (Cai, Thissen, & du Toit, 2016) software and modifications to the items and the instrument were made in accordance with the results of this analysis (e.g., revising/removing items with negative discriminations, unacceptable item difficulty indices, and/or unsatisfactory Std. UMS statistics).

3.2.5.2 Substantive Validity

This category refers to the empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks. Additionally, there is the requirement that a theoretical framework must be established that supports the development of the construct map and operational definition from which the internal and external models must logically follow. The data obtained from the cognitive labs and the usability study addressed substantive validity in this regard.

P-value fluency, as defined by this research, is the ability to use *p*-values in a manner that is methodologically defensible and does not “contribute to the statistical distortion of the scientific literature” (as defined by Greenland, et al., 2016). A researcher with high *p*-value fluency is one who can appropriately interpret *p*-values, both in and out of context, and recognize when others have not. Inspired by the framework of Greenland had his colleagues (2016), this paper proposes a construct map (Figure 3.1) that identifies level of proficiency based on a respondent’s ability to differentiate between misinterpretations and correct reporting of *p*-values.

Consistent with the construct map, this instrument was designed to assess the respondent at two process levels with the use of items that measured misinterpretations in a context-free setting and with items embedded in a research vignette setting. It is the speculative position of this research that the relative difficulty of the items is dependent upon the location within/without context, and that the former represents the more difficult task. Item hierarchy analysis of the field test data was used to investigate the strength of correlation between the item difficulty indices and the corresponding proposed process levels. These results informed the decision as to whether revision of the items or revision of the blueprint was the more prudent course of action.

3.2.5.3 Structural Validity

This category refers to the degree to which relationships between items conform to the theoretical view of the construct. The internal model for the construct specifies *p*-value fluency to include proficiency across two sub-categories of *p*-value misinterpretations (single *p*-values, comparisons and predictions) as exhibited in both context-free items and when evaluating the research of others. This structure is reflected in the instrument as the test specifications address both facets of the research process and all sub-categories of misinterpretations (see the test blueprint for details).

As stipulated previously, unidimensionality of the construct was assumed until contradictory evidence presented itself; however, the internal structure of the model suggested potential multidimensionality (two sub-categories of misinterpretations measured in two different settings). Exploratory factor analysis, therefore, was performed to check for consistency with the internal structure suggested by the data, with an eye to potentially eliminating items that did not load onto the dominant factor. Correlational analyses were also warranted to look at the relationship between each component and the overall construct.

As suggested by the internal model (Figure 3.3, Figure 3.4), a scholar must possess research skills in two main areas: competency in evaluating the research of others and competency in performing his own research (Schuyten, 1991). *P*-value fluency is demonstrated through competencies in both of these areas. The proposed instrument has a clearly defined purpose: to measure *p*-value fluency in PhD candidates, and has a defensible rationale for doing so: as a screening mechanism for future academicians to determine if they possess the requisite skill set to participate as a fully-functioning member of the research community, in other words, a researcher who will not "contribute to the statistical distortion of the scientific literature" (as defined by Greenland, et al, 2016). For this reason, the instrument is necessarily criterion-referenced.

3.2.5.4 Generalizability

This category refers to the degree to which test score properties and interpretations generalize to and across population groups, settings, and tasks including validity generalization of test criterion relationships. The target population for this instrument is future researchers: ideally, PhD candidates (outside the field of Statistics), who have completed their methodological training. It was the intention of the sampling process to capture a broad range of student characteristics to be inclusive of all genders, races, and programs of study. The demographic information was compiled to facilitate performance comparisons by subgroup and to allow Differential Item Function (DIF) analysis to be performed.

The overall reliability of the instrument was measured by the Coefficient Alpha and with Rasch/IRT model reliability estimates. These estimates are affected both by length of instrument and size/nature of the sample; thus, these statistics were examined with an eye for modifications therein. Based on the results of the EFA, the reliability estimates were used to assess the unidimensionality assumption and were computed for each sub-score, where indicated.

Secondary analysis was performed using Item-deleted reliability measures. Ideally, these values should be lower than the overall reliability. Measures that are higher than the overall coefficient indicate that the instrument would function better if that item were removed. Items that demonstrated cause for concern were reviewed and removed if indicated.

3.2.5.5 External Validity

This category refers to how the construct is expected to relate to other constructs and variables. When the field test was conducted, subgroup comparisons were investigated. Due to the varied nature of program requirements amongst PhD fields, there was a reasonable expectation that groups with differing levels of statistical preparation would perform disproportionately on this instrument. Tangentially related, owing to high survey response attrition, a thorough analysis of the characteristics of the persisters was justified.

The external model (Figure 3.2) situates statistical proficiency as a necessary component of PhD scholarship; therefore, it stands to reason that there is a continuum of proficiency that is navigated as one progresses through a plan of study. Longitudinal analysis, or group comparisons, should investigate this notion that performance will improve over time and/or that upper-division students (4th or 5th year) will outperform incoming students.

3.2.5.6 Consequential Validity

This category refers to the value implications of score interpretation as well as the actual and potential consequences of test use, especially regarding fairness. In the technical review of the items, the language was examined for use of offensive or stereotypic language. The items were written in a manner that rendered them devoid of ethnically-identifying pseudonyms and gender-specific pronouns, or at the very least, distributes them fairly. As mentioned previously, DIF analysis was performed on the field test data to identify any items with questionable discriminatory functionality.

The purpose of this instrument is to measure graduate students' proficiency in aspects of a statistical disposition most pertinent to their careers; e.g., communicating statistical results to others and contradicting claims made without proper statistical foundation. The results of the assessment will (eventually) discriminate between those students who are well-prepared to be fully-functioning members of the research community and those who need more preparation in this aspect of their plan of study. The use of this instrument as designed may result in stigmatization of certain groups (i.e., those unlikely to score well); however, further analysis is required in order to determine whether this is a source of invalidity (e.g. lack of access) or an inevitable consequence that accompanies any sort of high-stakes testing.

The intended application of this instrument is to identify students who possess a lack of statistical and research knowledge and to identify programs for which there is a shortage of adequate statistical preparation. In doing so, the information can be used to design remediation programs for individuals and to institute policy change at the department level in terms of plan of study requirements. It remains to be seen what the consequences of these actions might entail (individual student attrition, burden on department resources, etc.) and further analysis is required in this aspect of validity but is presently inestimable with only the pilot/field test data at hand and will not be part of this dissertation.

3.2.5.7 Summary of Validation Plan

The validation plan for this instrument was designed to be comprehensive in addressing the six components of validity proposed by Messick: Content, Substantive, Structural, Generalizability, External, and Consequential. Analysis conducted within each component was undertaken with the aim of improving individual item quality and the performance (and thus implications) of the instrument as a whole. Based on the results of the validation plan as outlined by the aforementioned analyses, the instrument was revised as indicated and a subset of items was selected to serve as the (current) final version of the instrument (PV3).

3.3 Phase IV: Instrument Administration and Data Analysis

The second research question, *What do the results of the KPVMI-1 administration tell us about the current level of p-value fluency among doctoral students nationally?,* was addressed via analysis of a subset of the field test data (n = 147) with respect to performance on the subset of items considered sufficiently validated as developed in Phases I-III (KPVMI-1). Inferences about respondents' *p*-value fluency were drawn from this data, in totality and by subgroup where appropriate.

3.3.1 Data Selection

The purpose of this research is to determine the extent to which future researchers struggle with interpreting and reporting *p*-values in the context of independent research and peer review. Having designed an instrument to measure respondents' *p*-value fluency, this phase of the research then utilized that instrument to obtain baseline data in a national sample. Greenland and his colleagues, in devising their list of misinterpretations, combed through existing research publications to identify errors that "statistically distort the scientific literature"; thus, this list is based on the contributions of *practicing researchers*. It is of interest in this study to determine if the next generation of researchers, i.e., current doctoral candidates, is as susceptible to these misinterpretations. The release of the ASA's Statement on *p*-values and other associated publications give reason to believe that perhaps the renewed attention this problem has been receiving might have begun to ameliorate the situation.

The data for this phase of the research was filtered from both field test samples. After merging both samples and removing cases based on ineligibility (Master's students removed) and response level (sparse and/or incomplete responses were eliminated when appropriate – a further discussion of this is presented in the Results section), the final dataset included 147 cases distributed across 16 institutions.

3.3.2 Data Analysis

Analysis of performance on the subset of items considered sufficiently validated as developed in Phases I-III (KPVMI-1) was performed to draw inferences about respondents' *p*-value fluency, in totality and by subgroup where appropriate. Initially, the data remained aggregated in order to generate a snapshot profile of *p*-value fluency nationally (i.e., to answer RQ2: *What do the results of the KPVMI-1 administration tell us about the current level of p-value fluency among doctoral students nationally?).* Appropriate descriptive statistics for overall performance, and by sub-category of misinterpretations where factor analysis indicated such a calculation was warranted, were performed. Secondary analysis compared performance figures by subgroup (i.e., by Major, Status, and Experience) and made inferences therein. Specifically, this phase of the research investigated the following sub research questions: *Are there performance differences by subgroup? Can these differences be attributed to differences in experience or preparation? What are the implications of these results in the context of the ASA's principles?* A thorough discussion of this analysis is presented in the Results section.

3.4 Summary of Instrument Development and Administration

In response to the first research question, the *KPVMI*-1 instrument, designed to measure *p*-value fluency, was constructed across three phases. Using the list generated by Greenland and his colleagues (2016) as the theoretical framework, the construct was mapped onto a continuum of fluency that describing a respondent's ability to differentiate between misinterpretations and correct reporting of *p*-values, both within the context of research vignettes and in context-free settings. The initial item pool was subjected to expert review to inform the preliminary test blueprint. A pilot test, consisting of cognitive labs and a usability study was used to revise the instrument and the blueprint in anticipation of the field test. The results from the field test administration were used to assess reliability and validity of the instrument. Based on the aforementioned analyses, a subset of items was identified to serve as the (current) final version of the instrument (PV3). Analysis of performance on the subset of items considered sufficiently validated as developed in Phases I-III (*KPVMI*-1) was performed to draw inferences about respondents' *p*-value fluency to answer the second research question and the subsidiary follow-up questions.

These methods employed to develop and validate the *KPVMI* (in pursuit of research question 1) were selected for their ability to reduce measurement error when the instrument is administered for research purposes (as in the objective of research question 2). The results of this process are discussed in the next chapter.

Chapter 4 – Results

In this chapter, the study results are reported as they pertain to each of the two research questions. Results from Phases I through III address the first research question, *Can a sufficiently reliable and valid measure of p-value misinterpretations (in a research context) be constructed*?. Specifics of each component of this process are presented: item development, pilot testing, and field testing. Results from Phase IV addresses the second research question, *What do the results of the KPVMI-1 administration tell us about the current level of p-value fluency among doctoral students nationally?.* Results of the baseline data analysis of the field test are presented.

4.1 Research Question 1

4.1.1 Phase I of Instrument Development

Using the proposed test blueprint as a guide (see Table 3), the first step in the instrument development process was to generate the initial item pool (PV0). The goal was to generate, at a minimum, four items for each of the 18 misinterpretations such that both correct and incorrect interpretations were captured in both a context-free setting and in the context of a research vignette. This was accomplished in phases with the True/False (context-free items) developed first, followed by the research vignettes.

4.1.1.1 Development of the Context-Free Items

In Greenland et al.'s original list, each statement is worded in the negative sense (i.e., as a misinterpretation and not a correct interpretation); thus, it was necessary to rewrite some in the affirmative in order to avoid having all T/F[10] items be correctly scored as "False." Negating

[10] At this point in the research, it was unclear which of the following dichotomies would be used: TRUE/FALSE or AGREE/DISAGREE or CORRECT/INCORRECT. Based on the wording of the original statement, one or another of the aforementioned choices might be more prudent than another. This issue was discussed and resolved by the item panel.

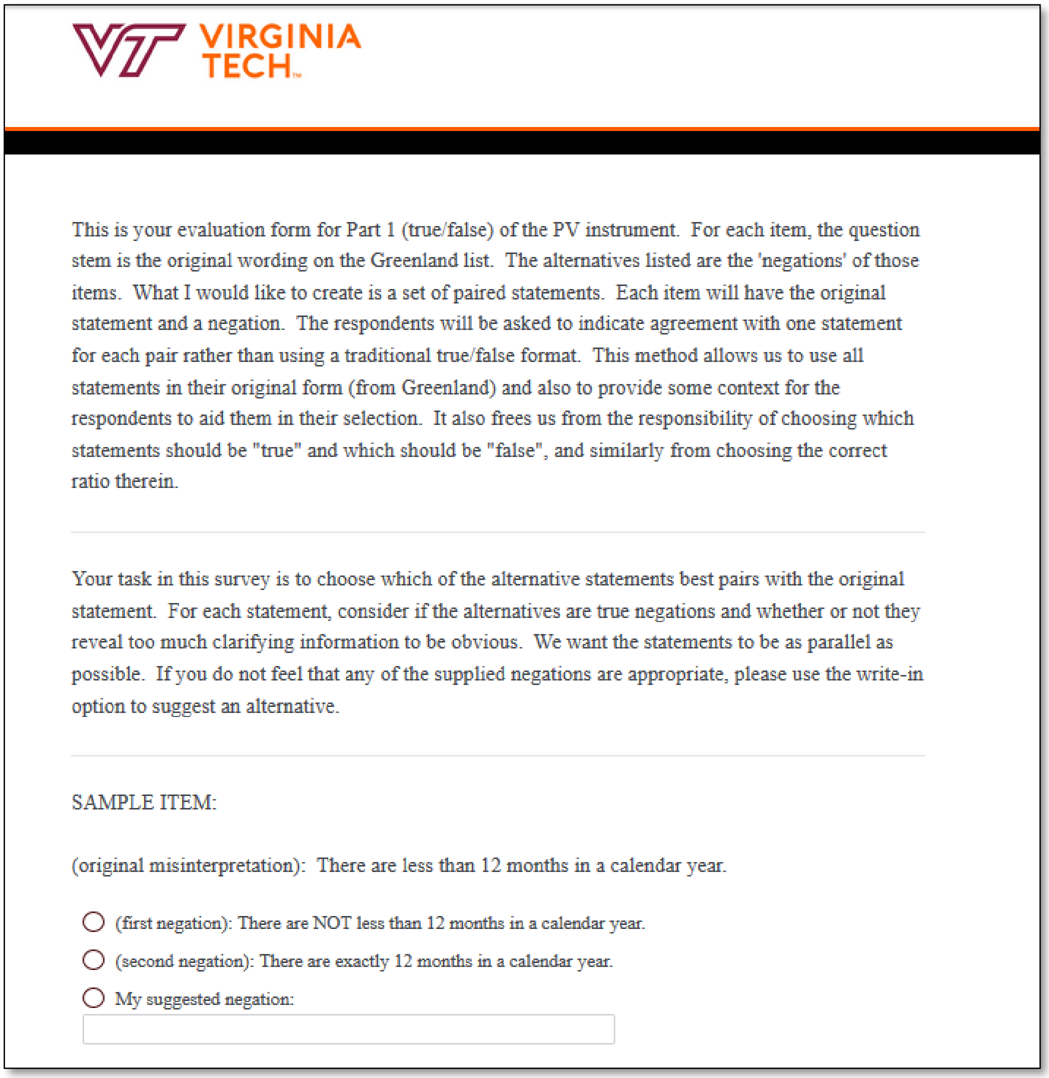
statements is often a complicated process that is not as simple as the mere insertion of the word "not." To that end, various negations were proposed to provide the item panel options from which to select the most appropriate. A sample item can be seen here:

The p-value is the probability that the test hypothesis is true.

- The p-value is not a measure of the probability of the truth of the test hypothesis.
- The p-value is NOT the probability that the test hypothesis is true.
- The p-value is not a hypothesis probability for any hypothesis.

In advance of the first item panel meeting, the panelists were sent a link to an electronic survey (administered in Qualtrics) that would gather feedback about the suitability of individual items in particular and overall feasibility of the items in general. The survey instructions can be seen in Figure 4.1 and the survey in its entirety can be seen in Appendix E.

The intention of this exercise was that the panelists would record their responses, the data would be tabulated for level of inter-rater agreement, and then this information would be presented at the item panel meeting in a forum that would allow for negotiation of disagreements and facilitate editing. In actuality, the presumption that this would be a straightforward task was naïve. A complete response was recorded for neither panelist and the consensus was that rather than using the item panel meeting to discuss the results, it was necessary to convene an item panel to determine what data collection should entail and to clarify misconceptions about the process.



This is your evaluation form for Part 1 (true/false) of the PV instrument. For each item, the question stem is the original wording on the Greenland list. The alternatives listed are the 'negations' of those items. What I would like to create is a set of paired statements. Each item will have the original statement and a negation. The respondents will be asked to indicate agreement with one statement for each pair rather than using a traditional true/false format. This method allows us to use all statements in their original form (from Greenland) and also to provide some context for the respondents to aid them in their selection. It also frees us from the responsibility of choosing which statements should be "true" and which should be "false", and similarly from choosing the correct ratio therein.

Your task in this survey is to choose which of the alternative statements best pairs with the original statement. For each statement, consider if the alternatives are true negations and whether or not they reveal too much clarifying information to be obvious. We want the statements to be as parallel as possible. If you do not feel that any of the supplied negations are appropriate, please use the write-in option to suggest an alternative.

SAMPLE ITEM:

(original misinterpretation): There are less than 12 months in a calendar year.

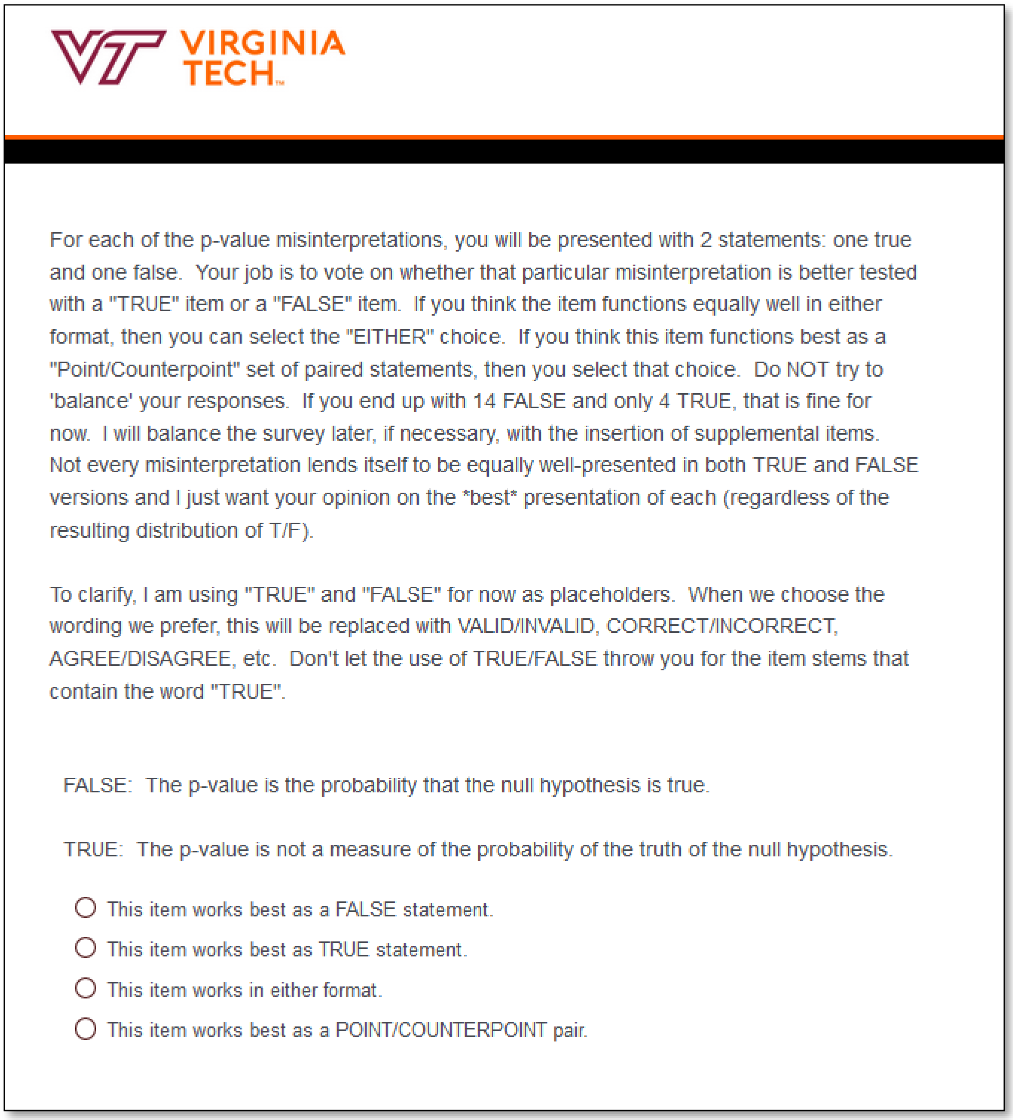
○ (first negation): There are NOT less than 12 months in a calendar year.
○ (second negation): There are exactly 12 months in a calendar year.
○ My suggested negation:

Figure 4.1. PV0(1) Survey Instructions

The first item panel meeting revealed many impediments to item development and data collection therein. The main difficulty the panel faced was in deciding the feasibility of each misinterpretation to function in both “True” and “False” presentations. While all statements work as “False” since the original list is one of *misinterpretations* (i.e., no statement on the list is correct/true), the panel’s sentiment was that not all statements work as “True” – this being problematic for the mere fact that it is not psychometrically sound to have a True/False test in which all the correct answers are “False.”

A secondary concern identified by the panelists was that the wording of many of the items was problematic – both in the researcher-generated item stems, but also as worded on the original list of misinterpretations (by Greenland, et al.). As an example, Greenland and his colleagues use the expression “test hypothesis” often in lieu of the more common “null hypothesis” and it was suggested by the panel that this turn of phrase would be confusing to respondents. Related to this issue of readability, many items were determined to be unnecessarily verbose and it was suggested that items be trimmed for brevity when appropriate. (Details of specific editing suggestions can be seen in Appendix F).

Minor language edits aside, the panel’s conclusion was that the most pressing concern at the moment was to determine the suitability of each misinterpretation to “True” and “False” formatting and pending those decisions, attempt to construct a balanced instrument. In advance of the second item panel meeting, the panelists were sent a link to an updated electronic survey (administered in Qualtrics) that would gather feedback about the suitability of individual items to particular formats. The survey instructions and a sample item can be seen in Figure 4.2 and the survey in its entirety can be seen in Appendix G. Note that in this version of the item pool, the items themselves remain relatively unaltered from the PV0(1) version of the survey save minor language edits identified by the panel, but that the *instructions to respondents* is what has changed.



For each of the p-value misinterpretations, you will be presented with 2 statements: one true and one false. Your job is to vote on whether that particular misinterpretation is better tested with a "TRUE" item or a "FALSE" item. If you think the item functions equally well in either format, then you can select the "EITHER" choice. If you think this item functions best as a "Point/Counterpoint" set of paired statements, then you select that choice. Do NOT try to 'balance' your responses. If you end up with 14 FALSE and only 4 TRUE, that is fine for now. I will balance the survey later, if necessary, with the insertion of supplemental items. Not every misinterpretation lends itself to be equally well-presented in both TRUE and FALSE versions and I just want your opinion on the *best* presentation of each (regardless of the resulting distribution of T/F).

To clarify, I am using "TRUE" and "FALSE" for now as placeholders. When we choose the wording we prefer, this will be replaced with VALID/INVALID, CORRECT/INCORRECT, AGREE/DISAGREE, etc. Don't let the use of TRUE/FALSE throw you for the item stems that contain the word "TRUE".

FALSE: The p-value is the probability that the null hypothesis is true.

TRUE: The p-value is not a measure of the probability of the truth of the null hypothesis.

○ This item works best as a FALSE statement.
○ This item works best as TRUE statement.
○ This item works in either format.
○ This item works best as a POINT/COUNTERPOINT pair.

Figure 4.2. PV0(2) Survey Instructions

As with the first item panel survey, the intention of this exercise was that the panelists would record their responses, the data would be tabulated for level of inter-rater agreement, and then this information would be presented at the item panel meeting in a forum that would allow for negotiation of disagreements and facilitate editing. Unlike the first survey, this iteration produced meaningful results. Of the 18 item pairs, agreement was recorded for 6 of them. Discrepancies were resolved based on the severity. Mild disagreement occurred when one rater indicated a preference (TRUE, FALSE, or POINT/COUNTERPOINT) and the other rater did not (selecting the EITHER choice). In this situation, the item was categorized according to the stated preference. Severe disagreement occurred when the raters indicated opposing preferences. In this situation, the item was designated as a P/CP pair as this assignment allowed both versions of the item to be retained. See Table 4.1 for details.

Table 4.1 – Rater Discrepancy Reconciliation Matrix

Level of Agreement	**1st Rating**	**2nd Rating**	**Assignment**
Agreement	TRUE	TRUE	TRUE
	FALSE	FALSE	FALSE
	EITHER	EITHER	TRUE*
Mild Disagreement	TRUE	EITHER	TRUE
	FALSE	EITHER	FALSE
	P/CP	EITHER	P/CP
Disagreement	TRUE	FALSE	P/CP

*These items were assigned TRUE to balance the instrument.

At the conclusion of this reconciliation process, there were seven items designated as FALSE, seven items designated as P/CP, and four items designated as TRUE. A complete list of item pairs, including modifications from the original list of misinterpretations, and rater data can be seen in Appendix H. Due to the imbalance of TRUE/FALSE items, the item pool was expanded to include additional TRUE statements. A set of six potential item stems were written and presented to the item panel for evaluation with the intention of retaining three or four of them in the pilot version of the instrument.

At the third item panel meeting, the panelists were presented with the items and asked to rank them in order of quality and viability (1 = best, 6 = worst). Table 6 shows the item stems accompanied by the panelists' rankings. Ultimately, only the lowest scoring item was removed and the remaining five of six items were retained with the understanding that after the pilot testing, the lowest performing items would be subject to removal if appropriate for balance. At this stage of the instrument development process, the panel felt it wise to err on the side of having too many items (rather than not enough).

Table 4.2 – Supplemental Item Ranking Log

Extra True Statements for Balance	Comments	R-1	R-2	Rater Average	Decision
The p-value is not the probability that any hypothesis (null or alternative) is true.		5	6	5.5	Remove
A definitive conclusion (i.e., accepting a particular hypothesis) cannot be deduced from a single p-value, no matter how large.		6	3	4.5	Retain
Error rates refer to the long run frequency of rejecting the hypothesis across repeated testing and not to the chance of error for any particular instance.		3	4	3.5	Retain
It is possible for two studies of the same hypothesis to produce identical (or similar) p-values yet exhibit clearly different observed associations.		4	2	3	Retain
Large effects can be masked by noise in the data and fail to achieve statistical significance, particularly in the case of small studies.	Is the term "noise in the data" known?	2	5	3.5	Query pilot students about the phrase
It is possible for two studies of the same hypothesis to produce p-values on opposite sides of 0.05 despite being in perfect agreement (i.e., may show identical observed associations).	Fix "may show identical observed associations"	1	1	1	Fix and retain

The third item panel meeting concluded Part I of the initial item pool development and resulted in a set of 23 items: 7 FALSE, 9 TRUE, and 7 P/CP pairs.

4.1.1.2 Development of the Contextual Items

Part II of the initial item pool development was the creation of the research vignettes. The objective was to draft short (~1 paragraph) descriptions of research settings that would be accompanied by a series of statements (based on the 18 misinterpretations) meant to simulate potential researcher actions and conclusions (both correct and incorrect). The stated task of the respondent would be to identify which of the statements represented appropriate researcher conduct.

Eight vignette pairs were originally proposed that represented a variety of statistical tests (t-test, ANOVA, Chi-square) and covered a range of situations (1-sided hypothesis, 2-sided hypothesis, independent samples, dependent samples). The scenarios were paired to allow both with/without statistical significance to be demonstrated. Research blurbs and accompanying statements were written for seven of the eight proposed situations and presented to the item panel for feedback. Not all of the misinterpretations were able to be incorporated into every scenario, but as many viable statements as possible were attempted. Table 4.3 lists the statistical situations by type and shows the tabulation of misinterpretation by vignette pair.

Table 4.3 – Distribution of Statistical Scenarios across Research Vignettes

	Misinterpretation																			
Vignette	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	**Pair Total**	**Type**
1A	x	x		x		x			x										7	Chi-square
1B	x	x	x						x	x										2tailp
2A	x				x				x			x		x					8	1-sample t
2B	x		x						x	x		x		x				x		1tailp
3A	x	x	x						x	x	x	x				x	x	x	12	1-sample t
3B	x	x		x					x			x			x	x	x			2tailp
4A	x	x		x	x				x	x		x		x					10	2-sample ind.t
4B	x	x		x		x		x	x			x		x						1tailp
5A	x	x	x				x		x	x	x	x	x	x				x	15	2-sample dep.t
5B	x	x		x	x	x		x	x			x	x	x						1tailp
6A	x	x		x	x	x		x	x			x		x					13	2-sample ind.t
6B	x	x					x		x	x	x	x		x				x		1tailp
7A																				2-sample ind.t
7B																				2tailp
8A	x	x		x		x		x	x			x							13	ANOVA
8B	x	x	x				x		x	x	x	x	x					x		2tailp

The inspiration for the vignettes came from *Beginning Statistics* (2^{nd} ed., 2014) by Warren, Denley, and Atchley. Borrowing scenarios from textbook examples and homework exercises, the vignettes were crafted through modifications and/or elaborations therein. A sample vignette (with accompanying statements) is presented here and the initial vignette pool in its entirety can be found in Appendix I.

> Scenario 3B (1-sample, 2-sided hypothesis): A manufacturer is responsible for making barrels used to store crude oil, which are designed to hold exactly 55 gallons of oil. As part of a routine quality check, a factory manager randomly tests 27 barrels off the assembly line and finds the mean capacity is 54.7 gallons with a standard deviation of 0.8 gallons. The p-value for the appropriate test procedure was .062.
>
> H0: the average barrel capacity is 55 gallons
> Ha: the average barrel capacity is not equal to 55 gallons

#1: The probability that the average barrel capacity is 55 gallons (i.e., the null is true) is 6.2%.

#2: The probability that chance produced the observed association (i.e., a sample mean of 54.7 gal) is 6.2%.

#4: A nonsignificant test result (p = .062 > 0.05) means that the average barrel capacity is 55 gallons (i.e., the null is true).

#17: In a second random sample of 27 barrels (mean = 55.3 gallons, standard deviation = 0.8 gal), the p-value for the appropriate test procedure was p = .062. Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as confirmatory.

#17: In a second random sample of 27 barrels (mean = 55.3 gallons, standard deviation = 0.8 gal), the p-value for the appropriate test procedure was p = .062. Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as being in agreement.

#17: In a second random sample of 100 barrels (mean = 54.85 gal, standard deviation = 0.8 gal), the p-value for the appropriate test procedure was p = .0637. Since the same hypothesis was tested twice and the resulting p-values are nearly identical, the manufacturer should interpret these results as being in agreement.

#12: This p-value should properly be reported as .05 < p < .10 or p < .10 (as opposed to p = .062).

#9: The probability of observing the data we did (i.e., an average barrel capacity of 54.7 gallons), if the null were true (i.e., average barrel capacity is 55 gallons), is 6.2%.

#16: In a second random sample of 35 barrels (mean = 54.7 gal, standard deviation = 0.8 gal), the p-value for the appropriate test procedure was p = .033. Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the factory manager should interpret these results as conflicting.

#15: In a second random sample of 25 barrels (mean = 54.7 gal, standard deviation = 0.8 gal), the p-value for the appropriate test procedure was p = .073. Since the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager should conclude that the totality of evidence upholds the null hypothesis (i.e., the claim that the average barrel capacity is 55 gallons).

At the first vignette panel meeting, the panelists decided which of the eight scenario pairs warranted further consideration. Factors contributing to the decision included ubiquity of statistical test, universality of context, readability of the scenario, and representation of misinterpretations. Four vignette pairs were selected to continue to the next round of development (2AB, 3AB, 5AB, and 8AB).

In accordance with the test blueprint and instrument design, each of the 18 misinterpretations needed to be tested at least once in context. Based on the abundant representation of the four selected vignette pairs, each misinterpretation was being tested across multiple scenarios. While having extra items is usually beneficial from an instrument design standpoint, overall test length was an important concern that necessitated removing some of the redundancies. Using a 4-step iterative process, the statements were pruned to achieve single representation of each of the 18 misinterpretations across a two-vignette pair subset. Selection was based on sparsity with the least represented misinterpretations (3, 5, 6, 7, 8, 11, 13, 15, 16, 17, 18) being assigned first and working backwards to those misinterpretations most frequently represented (1, 2, 9, 12). Consideration was given to balance across the scenarios and for readability where the choice was not forced. The result of this process yielded an *Odd* set of vignette pairs (3AB, 5AB) and an *Even* set of vignette pairs (2AB, 8AB) – each of which represented the 18 misinterpretations exactly once.   Details can be seen in Table 4.4.

Table 4.4 – Distribution of Misinterpretations by Vignette

	Misinterpretation																			
Vignette	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Type	Tally
3A	x	🟨	x						x	🟨	🟨	🟨				🟨	x	🟨	1-sample t	4
3B	🟨	🟨		x					🟨			x			x	x	🟨		2tailp	4
5A	🟨	🟨	🟨				x		🟨	x	x	🟨	🟨	x				x	2-sample dep.t	5
5B	🟨	x		🟨	x	x		x	🟨			🟨	x	🟨					1tailp	5
2A	x			x	x				x			🟨		x					1-sample t	5
2B	🟨		🟨						🟨	x	x	🟨		🟨		x	x	x	1tailp	5
8A	🟨	🟨		🟨		x		x	🟨			x			x				ANOVA	4
8B	🟨	x	x				x		🟨	🟨	🟨	🟨	x					🟨	2tailp	4

**Yellow square indicate possible representation, X marks selection*

In developing the initial vignette pool, the accompanying statements were all written in the negative (i.e., in terms of a *mis*interpretation). At this stage, for the four selected scenarios, it was necessary to develop paired statements so that this section of the instrument would follow a Point/Counterpoint format. Each of the panelists was presented with an advance copy of the research scenarios and accompanying P/CP statement pairs for review prior to the second vignette panel meeting. At this time, edits were made in accordance with the panelists' feedback. Specific details can be found in Appendix J.

4.1.1.3 Item Development Summary and Revised Test Blueprint

The objective of Phase I of this research was to generate the first iteration of the *KPVMI* instrument. Starting with an initial item pool based on the internal model of the construct and the preliminary test blueprint, item stems and research vignettes were modified in accordance with the expert review of the item panelists (as described in the previous sections) and were consolidated into a preliminary version (PV1) suitable for pilot testing with actual subjects from the target population. The updated test blueprint is presented here (Table 4.5) and all items by section can be seen in Appendix K.

Table 4.5 – Revised Test Blueprint (PV1 version[11])

	Process Level				
Item Content	**Context-Free**			**In Context**	
	TRUE	FALSE	P/CP	Odd Vig	Even Vig
Misinterpretations of single P-values	7	6	4	14	14
MI1		1		1	1
MI2		1		1	1
MI3			1	1	1
MI4		1		1	1
MI5	1	1		1	1
MI6		1		1	1
MI7			1	1	1
MI8	1		1	1	1
MI9			1	1	1
MI10	1	1		1	1
MI11	1			1	1
MI12	1			1	1
MI13	1			1	1
MI14	1			1	1
Misinterpretations of P-value comparisons and predictions	2	1	3	4	4
MI15			1	1	1
MI16	1		1	1	1
MI17	1	1		1	1
MI18			1	1	1

4.1.2 Phase II of Instrument Development (Pilot Test)

Phase II of the study gathered empirical evidence in support of the outcome space by way of a small pilot test that was supplemented with cognitive interviews and a usability study. Before administering the instrument on a large scale, the pilot study was conducted as a means of providing an additional opportunity to revise the items and implementation procedures as necessary in order to obtain the most accurate and usable data. The present pilot study achieved

[11] *Note that there are two versions of this instrument (odd/even). Each version includes all context-free items (23) and either the odd or even vignette items (18) for a total test length of 41 items. Phase IIA participants did all items, Phase IIB participants did only one (odd or even) vignette set.*

these aims by targeting the same population as intended for the ultimate data collection but on a smaller scale.

The current pilot study was conducted in two phases, both of which utilized purposeful samples of VT students. The respondents, selected on a volunteer basis, were asked to participate in either a cognitive lab (Phase IIB, target sample size n = 5) or the usability study (Phase IIA, target sample size n = 10).

4.1.2.1 Phase IIA Overview

PhaseIIA, the usability study portion of the pilot study, consisted of the participants completing an online version of the instrument administered via Qualtrics and was utilized to estimate response time and determine challenges individuals faced when registering responses. A copy of the Qualtrics version of the instrument as presented to participants can be seen in Appendix L.

4.1.2.2 Phase IIA Participants

The usability study participants were recruited based on satisfactory completion of a capstone course in quantitative methods (EDRE 6606 and STAT 5616, in particular). The reason for this criterion was that these students would have been exposed to an appropriate minimum level of methodological training and enrollment in these courses is limited to graduate students outside the field of statistics. No other demographic criteria (e.g., race, sex, program of study, etc.) were stipulated in terms of suitability for participation.

Potential participants were recruited in a variety of formats (recruitment materials can be seen in Appendix A). Firstly, a purposeful sample of classmates and colleagues was solicited via email. Secondly, faculty members responsible for teaching targeted research methods courses were contacted and asked to disseminate recruitment materials to their students. Finally, the graduate school weekly listserv was employed to advertise the study to graduate students campus-wide. Of the seven students personally contacted, six responded and five agreed to participate. Of the five professors contacted, none replied and no participants were recruited via this method. The graduate school listerv advertisement generated 25 responses, of which 8 participants were accepted into the study. For the potential participants recruited using this method, participation was primarily accepted on a first-come, first-served basis (although some participants were excluded for cause[12]) and was terminated when the target sample size had been exceeded due to financial considerations. A tabulation of the recruitment efforts can be seen in Table 4.6.

Among the 13 completed surveys, there was a 4:9 split of students identifying as male and female, respectively. In terms of race/ethnicity, 8 students identified as White, 3 as Asian, and 2 as Other. With regard to program of study, there were 3 from Education, 1 from Forestry, 1 from Psychology, 1 from Public Affairs, 1 from Sociology, and 6 whose program was unlisted and selected 'Other'. Nine of the students were PhD students: there was one 1st year, three 2nd years, two 3rd years, two 4th years, one 5th year; the four remaining students were either Master's students or an otherwise unidentified status.

[12] Master's students, students with insufficient statistical coursework, or students for whom statistics was irrelevant to their program of study were disqualified for cause when identifiable. Details are explained in the table.

Table 4.6 – Summary of Phase IIA Recruitment

ID	Contact	Appropriate?	Reason	Result
1	personal email	yes		completed
2	personal email	yes		completed
3	personal email	maybe	Not post-prelim, but does know p-values	completed
4	personal email	maybe	Post-prelim?, but does know p-values	completed
5	personal email	maybe	PhD, post-prelim?	no response
6	personal email	yes		forward to colleagues
7	personal email	yes		completed
8	grad list	no	Stats not relevant	
9	grad list	maybe	PhD, post-prelim?	completed
10	grad list	no	Stats not relevant	
11	grad list	no	Master's student	
12	grad list	no	No stats experience	
13	grad list	maybe	Status unknown	completed
14	grad list	maybe	Status unknown	completed
15	grad list	maybe	Status unknown	completed
16	grad list	maybe	Status unknown	pending
17	grad list	maybe	Status unknown	completed
18	grad list	maybe	Status unknown	completed
19	grad list	no	Master's student	
20	grad list	maybe	Status unknown	no response
21	grad list	maybe	Status unknown	no response
22	grad list	maybe	Status unknown	waitlisted
23	grad list	no	master's student	
24	grad list	maybe	PhD, post-prelim?	completed
25	grad list	no	Master's student	
26	grad list	maybe	Not post-prelim	
27	grad list	maybe	PhD, post-prelim?	2b participant
28	grad list	yes		completed
29	grad list	no	1st year	
30	grad list	maybe		
31	grad list	maybe		
32	grad list	maybe		

4.1.2.3 Phase IIA Items Data

The purpose of this pilot test was not to analyze the performance of the respondents *on the items*, but to assess the performance *of the items themselves*. That being said, a cursory glance at respondent performance is warranted to verify that scores are not inappropriately high or low relative to expectation. In this instance, overall respondents performed moderately well on the instrument with total scores indicating respondents correctly answered about 2/3 of the items, on average – results that are not unsurprising considering the sample (i.e., self-selected individuals likely to have an interest and/or aptitude for *p*-value interpretation). Specific scoring details, in total and by sub-section, can be seen in Table 4.7 and in Figure 4.3.

Table 4.7 – Phase IIA Scoring Statistics

Descriptive Statistics					
Instrument Section	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
T/F Score	13	0.44	0.81	0.6538	0.14347
P/CP Score	13	0.00	1.00	0.6264	0.30044
Odd Vig Score	13	0.22	0.94	0.6325	0.21700
Even Vig Score	13	0.39	0.89	0.6410	0.17947
Total Score	13	0.36	0.90	0.6402	0.16781

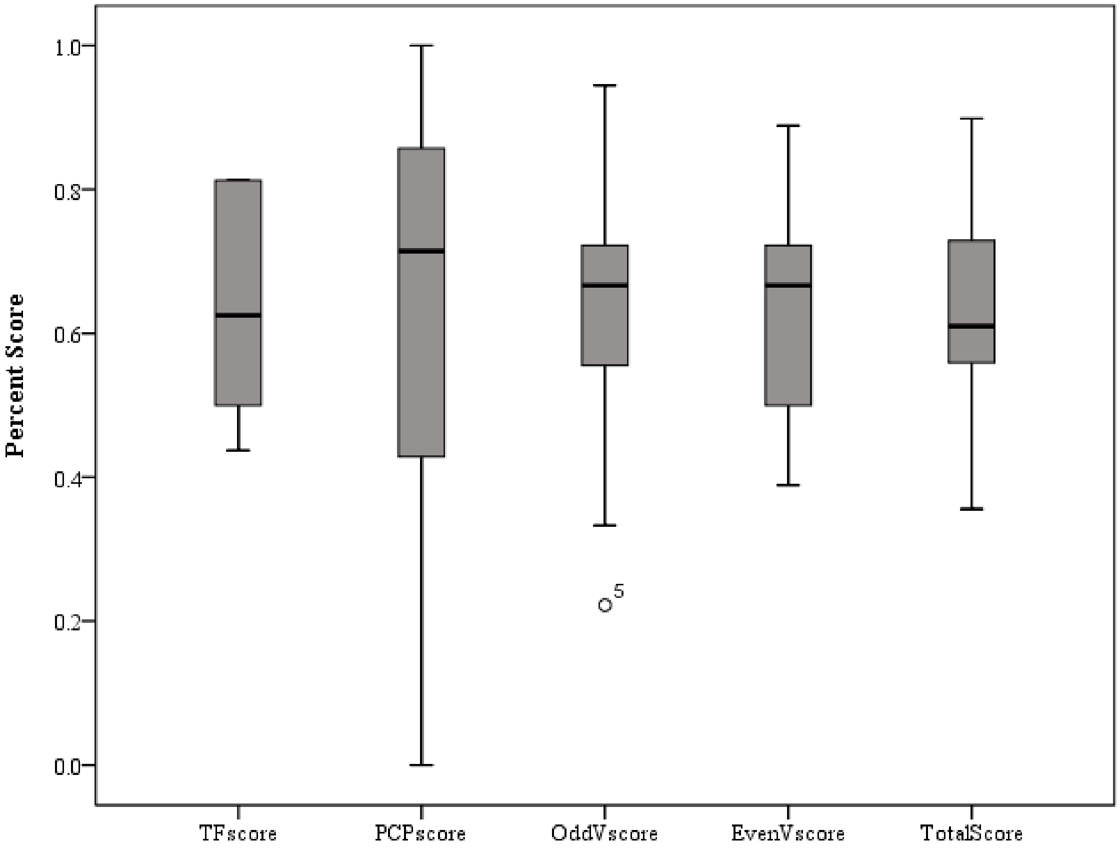
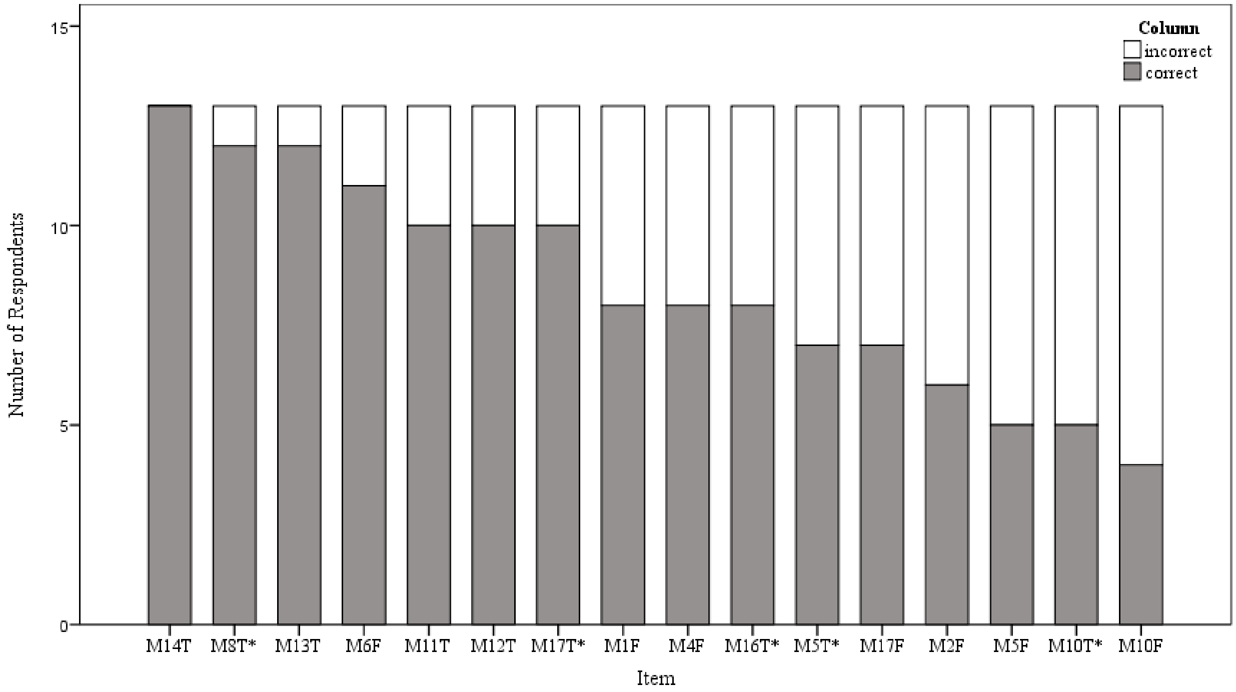


Figure 4.3. Phase IIA Distribution of Scores by Section

Turning our attention to the *item performance*, this will be addressed on the whole and via subsection. Considering the True/False section first, Figure 4.4 shows the percentage of correct answers by item arranged from least to most difficult. In determining whether or not the items are functioning as desired, one must take into consideration the scoring objective. In criterion-referenced assessments, the purpose is not necessarily in comparing respondents to each other, but in comparing scores to a reference threshold. In that situation, items that are answered correctly or incorrectly by a majority of the test-takers are permissible. However, in norm-referenced assessments, the purpose is to compare respondents to each other and items that do not discriminate between respondents at different levels offer little information. At this preliminary stage of the research, the presumption of a criterion-referenced application seems prudent; however, it is a bit premature to entirely rule out the possibility of standard setting in a future iteration. With that in mind, items that received more than 10 correct responses (n = 3) or less than 5 correct responses (n = 1) will be flagged for consideration, but will not be eliminated solely on this criterion[13]. Insight from PhaseIIB (cognitive interviews) will reveal whether these items are in and of themselves problematic or if in fact these items are correctly capturing the knowledge (or lack thereof) of specific identified misinterpretations.

[13] Items were identified as potentially problematic but not outright eliminated at this stage of validation because the small sample size of the pilot study did not justify removal on the basis of statistics alone.



*Additional (TRUE) items added for balance and do not represent original misinterpretations.

Figure 4.4. Participant Performance – True/False Section

A similar look at the Point/Counterpoint (P/CP) section reveals there to be only one potentially problematic item (for yielding 11/13 correct responses). The participant performance can be seen in Figure 4.5, ordered from least to most difficult.

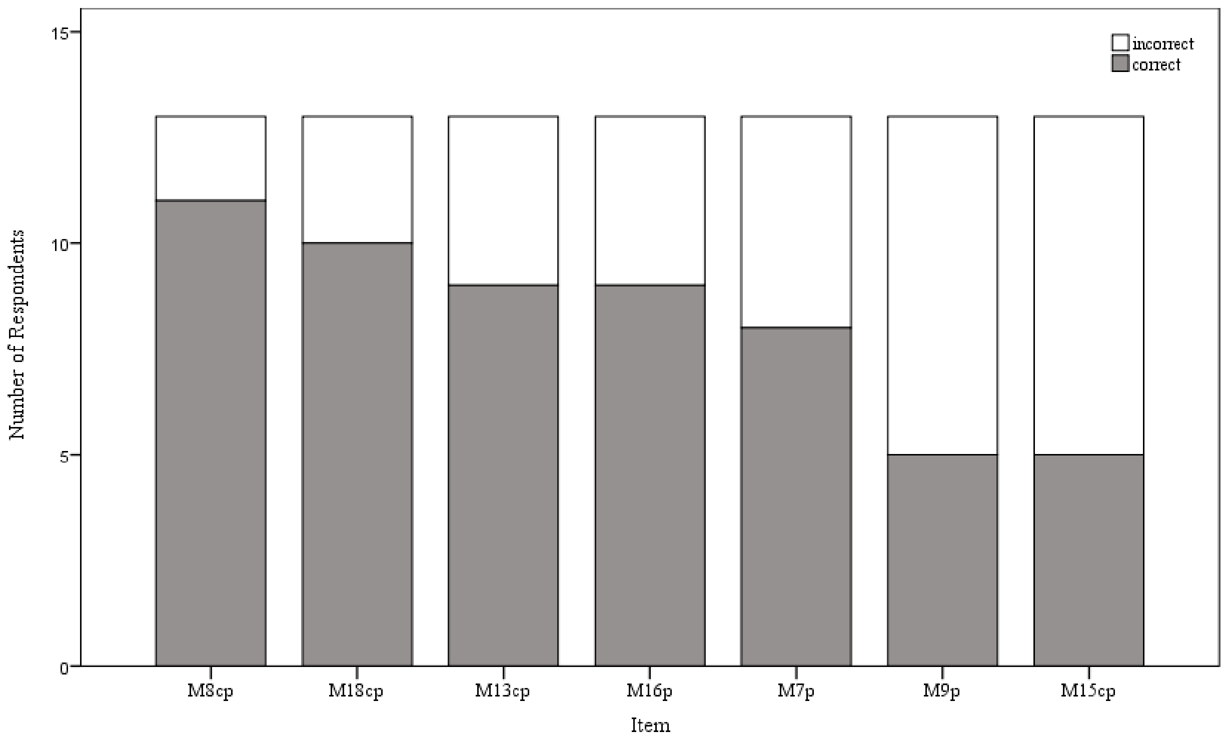


Figure 4.5. Participant Performance – P/CP Section

When considering the research vignette items, these were considered in odd/even subsets since each pairing captured each of the 18 misinterpretations exactly once. For both sets, there were no items that were flagged for being *too hard* (i.e., < than 5 correct responses), but there were a few flagged for potentially being *too easy* (i.e., > 10 correct responses); three items each for the even and odd sets (See Figures 4.6 and 4.7 for details, arranged in numerical order). Interestingly, however, it was not the same three misinterpretations that were flagged in each set (#12, #13, #14 for the odd set; #6, #14, #18 for the even set) raising the suspicion that the wording in one version was superior to the other since one would expect consistency (either evidence of a correct interpretation or evidence of a misinterpretation) across item formats. This is investigated further later in this research.

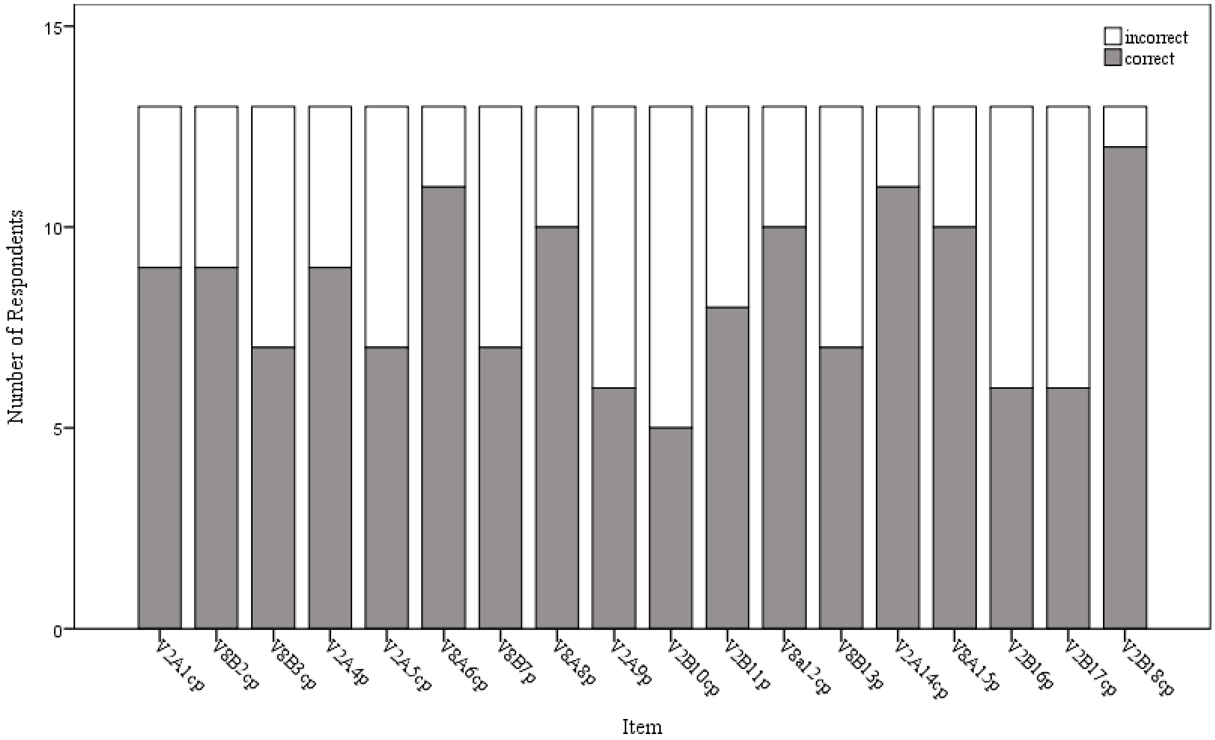


Figure 4.6. Participant Performance – Vignette Section (Evens)

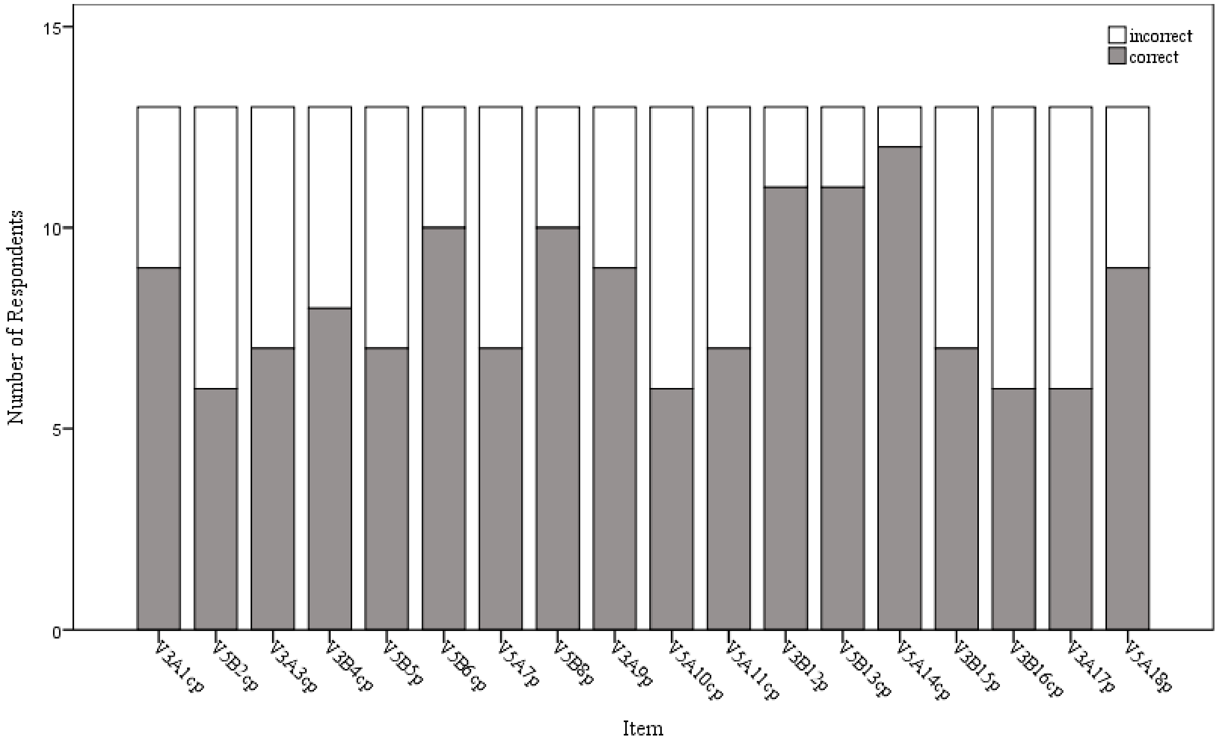


Figure 4.7. Participant Performance – Vignette Section (Odds)

In accordance with the proposed unidimensional construct, the internal consistency of the items across formats was investigated. Pearson correlations were computed to measure the strength of the relationship of each subsection score with the total score. In this assessment, high positive correlations (i.e., close to 1.0) were desired as this would indicate that higher scores on a given subsection correlates with higher scores overall. Negative correlations would indicate an inverse relationship between performance on the subsection and the instrument overall. Correlations low in absolute value (i.e., close to 0) would indicate a lack of correlation between the subsection and the overall score – in other words, performance on this section was not providing any useful information regarding the person's overall ability. Table 4.8 shows the correlation matrix (cells display *r* with the *p*-value in parentheses). The research vignettes showed the strongest correlation with total performance, and this was true even when considering the odd and even subsets separately. The True/False section showed the weakest relationship and was noticeably lower than the other correlations. This surprising finding is investigated further later in this research.

Table 4.8 – Correlation Matrix Phase IIA

Correlations					
	Total Score	T/F Score	P/CP Score	OddV Score	EvenV Score
Total Score					
T/F Score	.650 (.016)				
P/CP Score	.897 (< .001)	.547 (.053)			
OddV Score	.912 (< .001)	.453 (.120)	.701 (.008)		
EvenV Score	.916 (< .001)	.379 (.201)	.862 (< .001)	.808 (.001)	
Coursework	.698 (.008)	.299 (.320)	.842(< .001)	.557 (.048)	.703 (.007)

In addition to correlations between test subsections, the correlation between scores and coursework was investigated as a proxy for convergent validity. Theoretically, students who have completed more relevant coursework should have a better understanding of the statistical ideas and suffer from fewer misconceptions and misinterpretations. The total score correlated satisfactorily with this measure (*r* = .698, *p* = .008) with varying performance when considered by subsection. The P/CP (*r* = .842, *p* < .001) and Vignette sections (*r* = .557, *p* = .048; *r* = .703, *p* = .007) demonstrated moderately strong relationships; however, the T/F subsection continued to raise concern with a shockingly low *r* = .299 (*p* = .320).

Linear regression was employed to model the relationship between courses completed and total score. The suggested relationship ($\beta_0 = .421, SE_{\beta_0} = .076;\ \beta_1 = .081, SE_{\beta_1} = .025;$ ANOVA *F* = 10.428; *p* = .008) indicated that a person with no experience would score about 42% on average with an expected increase of about 8% for each additional course completed. See Figure 4.8 for details.

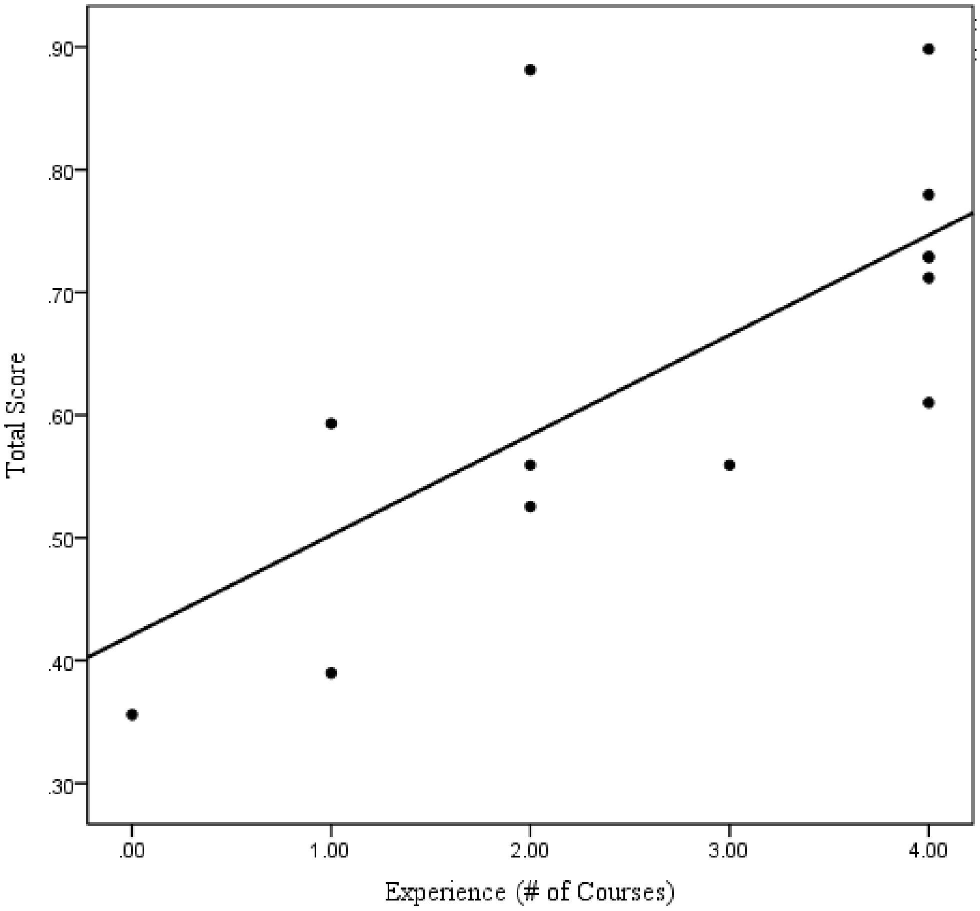


Figure 4.8 Regression Results – Phase IIA

The final assessment of instrument performance was a check on internal reliability, both in total and by subsection. When using Cronbach's alpha, the closer a value is to 1.0, the more reliable the measure. Conventional wisdom suggests that a value of 0.70 is the lowest acceptable threshold (Nunally, 1978). As can be seen in Table 4.9, this threshold has been met for the instrument as a whole and by subsection for the P/CP and the Odd vignettes. As seen with the other analyses, the T/F subsection had the poorest performance.

In diagnosing the reliability of a collection of items, an important follow-up analysis is to examine the *Corrected Item-Total Correlation* and/or *Cronbach's Alpha if Item Deleted* measures. These diagnostics will indicate which items, if any, are problematic and adversely affecting the instrument's reliability. In the case of *Cronbach's Alpha if Item Deleted,* values that exceed the alpha of the collection of items suggest that the instrument would be improved if these items were removed. In the case of *Corrected Item-Total Correlations*, negative values indicate that there is an inverse relationship between performance on this item and the total score. In this analysis, items were classified as 'poorly functioning' if the *Corrected Item-Total Correlations* fell below the 0.05 threshold. This criterion identified no issues with the P/CP items, a handful with the vignettes (n = 6 for even, n = 4 for odd), and quite a bit for the T/F items (n = 10/18). This is consistent with the other diagnostic measures that suggested that there was cause for concern with the T/F subsection.

Table 4.9 – Cronbach's Alpha Data Phase IIA[14]

Subset	# of Items	Alpha	Comments
T/F	16	0.42	10 items are poorly functioning
P/CP	7	0.754	
RV even	18	0.65	6 items are poorly functioning
RV odd	18	0.773	4 items are poorly functioning
All items	59	0.883	

[14] Full results of this analysis are in Appendix M.

4.1.2.4 Phase IIA Usability Data

The results generated from the exit interview were quite promising in terms of the respondent test-taking experience. Average response time was 54 minutes (standard deviation of 27.5 minutes). All participants indicated no issues with accessibility and 12 of the 13 participants found no issue with the survey navigation or font. Participants generally felt that the directions were easily understood (13/13 agreed for the T/F section, 13/13 agreed for the P/CP section, and 12/13 agreed for the Vignette section) and indicated a preference for the current formatting of the sections (12/13: T/F first, P/CP second, Vignettes last). There were no identified issues with the reading level of the instrument (13/13 agreement), nor generally with the arrangement of items per page (11/13 agreement). The lone source of disagreement pertained to the formatting of the vignettes. Only 6 of 13 respondents felt that the length of the scenario was appropriate, with 2 indicating it was too short and 5 indicating it was too wordy. Furthermore, the ratio of items per vignette was deemed appropriate by only 6 of 13 respondents, with 3 desiring fewer and 4 desiring more. In terms of incentivization, the results were inconclusive as to the best choice. 5/13 participants indicated that they would only participate in this type of research if being paid. For the remaining 8 participants who were willing to consider alternative inducements, all options were basically equally attractive: Extra Credit (6/8), Homework (7/8), Raffle (6/8); surprisingly, 7 of 8 participants indicated they would be willing to participate even in the absence of any incentive.

4.1.2.5 Phase IIB Overview

In PhaseIIB, cognitive labs, also referred to as “think-aloud” interviews, were used to gather additional empirical evidence in support of the outcome space. In this investigation, a small, but representative, sub-sample of the target population (n = 5 for the present study) was asked to complete the assessment while actively describing their thought process as they responded. The cognitive lab participants completed the same version of the instrument as the usability study participants (the online Qualtrics version[15]), but were video and audio recorded while doing so. The participants were encouraged to "think-aloud" as they answered the items. This commentary provided insight into the participant's level of statistical reasoning while also supplying feedback regarding aspects of the user experience. The former was used to determine if the instrument was able to distinguish between respondents at varying ability levels, while the latter was used to improve the usability of the instrument itself. The recorded sessions were transcribed for later analysis, selected details of which can be seen later in this section and in Appendix N.

4.1.2.6 Phase IIB Participants

Cognitive lab participants were recruited based on selection criteria as explained in Section 3.2.3.2 of this paper. Five potential participants were contacted via email and all five targeted individuals agreed to participate. Incidentally, the Phase IIA recruitment efforts yielded two additional potential Phase IIB participants; however, one declined to participate and the other was eliminated due to financial considerations.

Among the five respondents, there was a 4:1 split of students identifying as female and male, respectively. In terms of race/ethnicity, four students identified as White and one as Black. With regard to program of study, there was 1 from Mathematics, 1 from Mathematics Education, 1 from Educational Research and Evaluation, and 2 from Higher Education. All five participants were PhD students in either their third or fourth year of study (1 and 4, respectively).

[15] Four of the five participants were recorded taking the assessment online via Qualtrics to mimic an authentic user experience. The fifth participant was unable to do so due to technical issues and was recorded while completing a paper-based version of the instrument.

4.1.2.7 Phase IIB Items Data

Presentation of the Phase IIA Items Data results began with a cursory glance at respondent performance to verify that scores were not inappropriately high or low relative to expectation. That will be repeated here for Phase IIB. In this instance, overall respondent performance on the instrument was slightly higher than in Phase IIA with total scores indicating respondents correctly answered about 3/4 of the items, on average (compared with 2/3 in IIA) – results that are not unsurprising considering the sample. The credentials of the Phase IIB participants were somewhat superior to those of Phase IIA (specifically with regard to degree progress and methods courses completed) and thus the higher scores meet with expectation. Specific scoring details, in total and by sub-section, can be seen in Table 4.10 and in Figure 4.9.

Table 4.10 – Phase IIB Scoring Statistics

Descriptive Statistics					
Instrument Section	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
T/F Score	5	0.63	0.94	0.7502	0.11706
P/CP Score	5	0.57	1.00	0.7712	0.19185
RV Score	5	0.61	0.94	0.7776	0.13584
Total Score	5	0.61	0.95	0.7659	0.12975

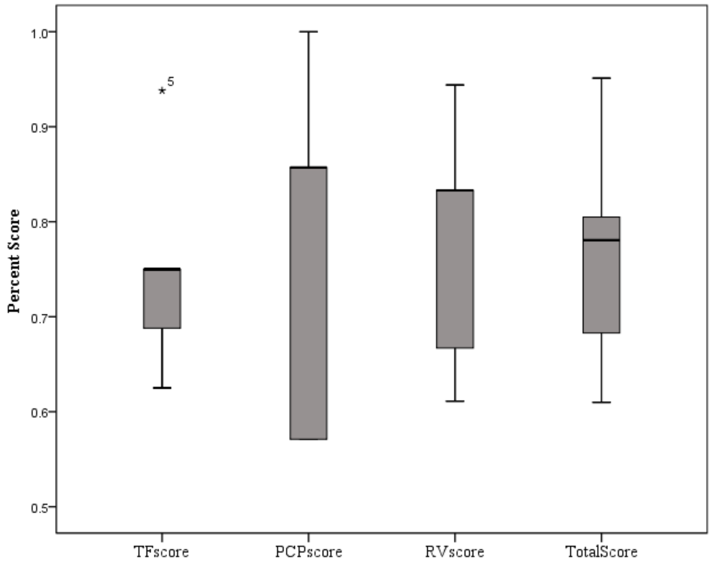


Figure 4.9 Phase IIB Distribution of Scores by Section

For the sake of parallelism, there were two[16] analyses from Phase IIA that were replicated with the Phase IIB data. Firstly, the internal consistency of the items across formats was investigated. Pearson correlations were computed to measure the strength of the relationship of each subsection score with the total score. In this assessment, high positive correlations (i.e., close to 1.0) were desired as this would indicate that higher scores on a given subsection correlates with higher scores overall. Table 4.11 shows the correlation matrix (cells display *r* with the *p*-value in parentheses). The research vignettes showed the strongest correlation with total performance. The True/False section showed the weakest relationship and was noticeably lower than the other correlations. These results are consistent with those seen in Phase IIA.

Table 4.11 – Correlation Matrix Phase IIB

Correlations				
	Total Score	T/F Score	P/CP Score	RV Score
Total Score				
T/F Score	.880 (.049)			
P/CP	.953 (.012)	.698 (.190)		
RV score	.979 (.004)	.765 (.132)	.989 (.001)	
Experience	.841 (.074)	.489 (.403)	.953 (.012)	.931 (.021)

The second analysis repeated from Phase IIA was an investigation of the correlation between scores and coursework as a proxy for convergent validity. Theoretically, students who have completed more relevant coursework should have a better understanding of the statistical ideas and suffer from fewer misconceptions and misinterpretations. The total score correlated satisfactorily so with this measure (r = .841, p = .074) with varying performance when considered by subsection. The P/CP (r = .953, p =.012) and Vignette sections (r = .931, p = .021) demonstrated moderately strong relationships; however, the T/F subsection continued to

[16] Not all analyses were repeated as the small sample size and data collection procedure were incompatible for such measures (e.g., item-by-item analysis of “hard”/”easy” items, and internal reliability by sub-scale).

raise concern with a dismally low r = .489 (p = .403). All of these relationships are consistent with the Phase IIA results, but the lack of statistical significance and low sample size require that these results be considered tentative.

Linear regression was employed to model the relationship between courses completed and total score. The suggested relationship indicated that persons with four completed courses would expect to score approximately 20% higher on average than those with three completed courses ($\beta_0 = .049, SE_{\beta_0} = .269$; $\beta_1 = .199, SE_{\beta_1} = .074$; ANOVA F = 7.239; p = .074). See Figure 4.10 for details. The utility of this analysis is fairly limited as the narrow range of enrolled courses renders our interpretation of the y-intercept meaningless and only really provides insight into how the completion of a fourth methods course improves performance (unlike in Phase IIA where the per course margin was computed); however, it is important not to discount this contribution to instrument validation. As expected, increased coursework leads to improved performance. If this analysis had indicated the reverse to be true, this would suggest the instrument to be flawed.

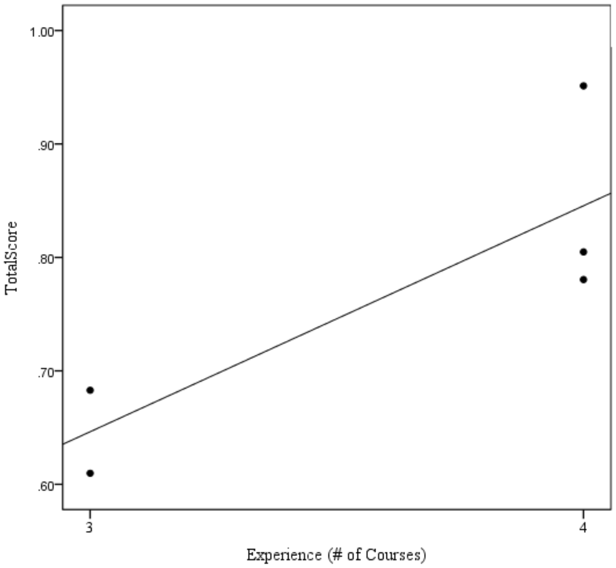


Figure 4.10 Regression Results – Phase IIB

4.1.2.8 Phase IIB Items Data by Participant

While the overarching objective of this phase of the research remains item and instrument development, the primary aim of the cognitive labs was to determine if the instrument was able to distinguish between respondents of varying ability levels. In Phase IIA, there was no way to corroborate scores on the instrument with the participants' true *p*-value conceptions and fluency due to the nature of the testing format and the absence of supplemental data (i.e., course grades, convergent measures). This was not so in Phase IIB where the depth of data provided by the think-aloud interview afforded the opportunity to not only determine the depth of *p*-value understanding of the participants but also supplied a ranking mechanism (albeit a subjective one) used to gauge the consistency of performance on the assessment with fluency.

At the conclusion of each session, the recordings were selectively transcribed and analyzed. The objective of this data collection was not to record verbatim every utterance, but rather to mine the session for rich excerpts that would provide evidence to inform the participants' *p*-value knowledge profile or that would reveal flaws in the instrument. A running record, by item and by participant, was kept that tabulated identified issues in the item wording, evidence of (mis)understanding, disposition of response (correct/incorrect), and where applicable, determination of fluency (an attempt to discern, based on the comments provided, if the participant truly understood the underlying idea[17]). A *p*-value profile for each participant, accompanied by sample excerpts, is presented here. The complete video transcript log can be seen in Appendix M.

[17] The motivation for the latter of these two categories was that not all problematic items are readily identified as such by participants in the moment; specifically, if a participant answers incorrectly, despite possessing the requisite knowledge, this reveals a flaw in the item necessitating consideration for revision or removal. Thus, when possible, item dispositions were cross-referenced with fluency as an additional measure of substantive validity.

The first participant profiled was Abigail[18]. She was judged as having a moderate understanding of quantitative methods in general and *p*-values in particular. Presented here are a few sample transcript excerpts[19]:

- No, they have not proved there to be no association…well, demonstrated, bother…[I don't want you to second-guess, you picked the correct answer. Do I need to change the word *demonstrated*?] well, demonstrated seems like evidence of, so…in my head I was thinking demonstrated versus proved [I think I was using demonstrated as a synonym for proved]. So, I originally read it as proved and was going to choose and then I was backing off as well what does demonstrated mean? Those aren't quite the same in my head.

- So, we are never allowed to accept the null hypothesis, that is just not good...in terms of if we are getting a giant p-value, then it is possible that we are thinking about this wrong and that there is actually more validity or it is actually true or it is possible that the null hypothesis is true - we could not conclude that it is definitely true but I guess it would be evidence in favor of thinking about it but you still could not come to a firm conclusion so I guess I would agree… about accepting the null, from what I heard having taken other stats classes, is the immediate, 'no, you should never say that'

- Yeah, I mean if you are talking about if something is equal/not, then you need a 2-sided test, but if you have very good evidence that it is >,< you should use a one-sided test…[does the *p* have to match the hypothesis?] I don't know what that means…still not sure what that means, if you are using a 1-sided test, your 5% rejection region is going to allow you to have a much closer to the middle critical value, so it will change what your critical value is...[do you know how to calculate a 2-tailed *p*-value?] I know how to calculate a critical value. [more clarification] I know there is a formula for calculating *p*-values, but I don't have it memorized, it is related to like you have your mean, depending on direction you have your +/-, then you have *z*, *t* and it is related to the number of standard deviations and it should be standard deviations over root n...I still don't understand...so when you are calculating your *p*-value, you are going to be choosing which *z* or *t* value according to if it is 1-sided or 2-sided, is that what you mean?

She appears to have a firm grasp on some main concepts, but it is a bit difficult to discern if this is based on a rich understanding of the ideas or a masterful recitation of textbook bullet points (e.g. “never accept the null hypothesis”). It is almost as if she knows just enough to be dangerous - in other words, she possesses a passing familiarity with ideas peripheral to *p*-values

[18] All participant identifiers were self-selected pseudonyms.

[19] In all transcript excerpts, the use of […] is used to denote the researcher comments. All other text is the utterances of the participant.

(sample size, effect size, Type I error) and statistics in general (sampling error, experimental design), but has not yet fully reified the relationships therein.

She was very concerned with exact/literal meanings of specific words/phrases, often to her detriment. In particular, she seemed preoccupied with the idea of *proving* a hypothesis and whether or not the verbs used in the stems were synonymous with *prove*. Even on concepts that she seemed to understand, her preoccupation with the lexicon interfered with her ability to answer (almost like she talked herself out of the right answer or convinced herself that we had reached a limit of her knowledge when perhaps we had not). Her ability to differentiate between a statistical decision (reject/not reject) and the corresponding conclusion was fleeting and inconsistent in that while able to envision them as distinct, her answers tended not to reflect this position. When she thought she was unsure/incorrect, she was not shy about asking for and considering clarification; however, I think at times her tendency to doubt her level of knowledge led her to accept scaffolding that was perhaps unnecessary and unfairly prejudiced the assessment of her ability.

In some cases, she clearly held one of the tested misconceptions and the instrument was able to capture that. More importantly, the instrument was consistently able to diagnose this misinterpretation across item formats (e.g., Misinterpretation #10 was answered incorrectly in all presentations: TRUE, FALSE, and Vignette formats; Misinterpretation #11 was answered incorrectly in both presentations: TRUE and Vignette formats; Misinterpretation #9 was answered incorrectly in both presentations: P/CP and Vignette formats). In the cases where she was unable to choose an answer, she seemed inclined to assume that the item itself was flawed rather than acknowledging her own lack of understanding. Providing further clarification had mixed results. In the cases where it led to a correct answer, the conclusion was normally that a single problematic word or phrase was hindering her selection process and suggested that a minor edit would render this item suitable for future participants. In those cases where the clarification did not lead to a correct answer, she was willing to accept that this must be a limit of her knowledge and not necessarily a problematic item. Either way, her performance seems to be evidence in favor of the instrument's ability to reflect participant ability.

The next participant profiled was Babs. She was judged as having moderate to high understanding of quantitative methods in general and *p*-values in particular. Presented here are a few sample transcript excerpts:

- Ok? I think I know what that means. Ok, I don't know the exact value of *p* from this statement, so just from that, I know that *p* could be any value < .05, not that *p* is actually many values, it is just that I am not reporting the exact value…which is just a mathematical statement...I can do away with all the rest of this sentence because these are just mathematical statements that don't mean the same thing at all.

- I think you are getting at certainty here, can you be certain because you have that magic low number?

- not sure what you mean by *perfect agreement*…are you telling me that the actual data points are identical? Or are you telling me that the studies are done on similar populations? That question is confusing me. If I stop and think, I think I can answer it. Yes, because it depends on the variation within each data set. I think that *agreement* is a very general term and it could mean many things and you have given me one example of what that means but you haven't said that's the only thing.

- So by definitive conclusion, are you saying that you are certain that your hypothesis is true or are you accepting that hypothesis? [sometimes the decision and conclusion mean the same thing, sometimes they don't] I think my trouble is…by definitive, you can be certain in your conclusion, like absolute truth? [If the *p* were very tiny, would that prove the null were false?] No, absolutely not...So, your wording is a *definitive conclusion*, does *definitive conclusion* mean I have absolutely no doubt about what I am saying, absolute truth...No matter how large the *p*-value, one cannot be certain about one's conclusion (for instance, accepting a particular hypothesis)

She appears to suffer from a misguided understanding of the definition of *p*-value, but is able to discuss related ideas fluently and understands the interplay between the components of the test statistics (e.g. sample size, variation, etc.).

In many cases, her decision-making was limited by her preoccupation with *absolute truth* with regard to hypotheses and the *magic number* heuristic of p-values. In both cases, her thinking was properly aligned; however, she tended to over-simplify and miss the nuance in some of the items because a large portion of them seemed to her to be able to be reduced to one of these two ideas (e.g., MI5T, MI15cp, MI16p, V3B16, V3A3, V3A17) . Like Abigail, she too was very concerned with exact/literal meanings of specific words/phrases, often to her detriment. Even on concepts that she seemed to understand, her preoccupation with the lexicon interfered with her ability to answer (almost like she talked herself out of the right answer or convinced herself that we had reached a limit of her knowledge when perhaps we had not). Her ability to differentiate between a statistical decision (reject/not reject) and the corresponding conclusion was fleeting and inconsistent in that while able to envision them as distinct, her answers tended not to reflect this position. When she thought she was unsure/incorrect, she was not shy about asking for and considering clarification. When she answered confidently without assistance (declaring "I understand") and was incorrect, I believe this is evidence of a heuristic she firmly holds (and thus the instrument is correctly capturing her misinterpretation). In other words, that false statement resonates with something she believes.

In some cases, she clearly held one of the tested misconceptions and the instrument was able to capture that (e.g., Misinterpretation #17 was answered incorrectly in two of three presentations: FALSE and Vignette formats; Misinterpretation #9 was answered incorrectly in both presentations: P/CP and Vignette formats). In other cases, after being given a brief explanation/tutorial of the concept, she was able to choose the correct answer (e.g., MI15cp, MI10cp) implying that a person who already possessed this knowledge should be able to recognize the correct selection for the items. In either case, her performance seems to be evidence in favor of the instrument's ability to reflect participant ability.

The next participant profiled was Dana. This participant was judged as having fairly weak understanding of quantitative methods in general and *p*-values in particular. Presented here are a few sample transcript excerpts:

- I understood the CP…so because I understood that one, I am going to choose P. [Did you not understand P?] Not as well, but not because it was not written well. I think I am getting tripped up between like the null hypothesis and my understanding, a lack of remembrance [sic].
- I am forgetting what the null hypothesis is, but I do understand the question because you are basically asking me what the definition of a *p*-value is.
- I understand both of these but I am not certain about it, I am just going to choose P
- I know what a *p*-value is and I know what a Type I error is…the *p*-value is the probability that you got your result due to random chance…I am disagreeing with that because a .05 *p*-value is not a reason to reject because of a Type I error
- Not sure I understand this, I kind of get it, but I don't…No, they should not be reported as exact values because anytime I have read a research paper, they have been reported as $p < .05$… There are 3 ways to write it with stars…

She is able to recognize *buzz words* but does not reason fluently about the underlying concepts (e.g., V5B8). In many cases, her decision-making was limited to the *lesser of 2 evils* approach: if she only understood one of the two options, that was her basis for selection (e.g., MI3cp, MI16p, V3A3cp, V3B15, V5A10, V5B5, V5B6, V5B13). She seems to be conflating agreement with truth – if she understands the statement, she is inclined to agree or think it must be true and if she doesn't, she tends to think it is false.

In some cases, she clearly held one of the tested misconceptions and the instrument was able to capture that (e.g., Misinterpretation #2). In other cases, after being given a brief explanation/tutorial of the concept, she was able to choose the correct answer implying that a person who already possessed this knowledge should be able to recognize the correct selection for the items. In either case, her performance seems to be evidence in favor of the instrument's ability to reflect participant ability. Her overall score is slightly better than 50% (slightly above probabilistic expectation) which would be consistent with a true understanding of some of the ideas.

Like Babs, when she thought she was unsure/incorrect, she was not shy about asking for and considering clarification. When she answered confidently without assistance (declaring “I understand”) and was incorrect, I believe this is evidence of a heuristic she firmly holds (and thus the instrument is correctly capturing her misinterpretation). In other words, that false statement resonates with something she believes.

The next participant profiled was Mike. He was judged as having moderate to high understanding of quantitative methods in general and *p*-values in particular. Presented here are a few sample transcript excerpts:

- I am disagreeing because of the word *perfect.*
- That is one of my favorite ones, that something can be statistically significant, but not practically important.
- I agree it is confirmatory but very rare to say yes for sure; you did this twice, how sure are you that this thing is actually true? I am reading confirmatory as, yes this confirms things, it moves me further towards believing it.

He is able to discuss ideas fluently and make commentary on mistakes he has observed others making, including the difference between *best practice* and *convention* (e.g., MI7p, MI12T). In many cases, his decision-making was limited by *over-thinking*, i.e., reading into the semantic nature of some words and almost trying to *game the test* as opposed to just allowing his natural conceptions to be demonstrated. The T/F specifically seemed to be problematic because he was looking for the key word or phrase that was making the statement false (looking for tricks?). Even when able to reasonably argue for/against the statement, he still seemed hesitant to pull the trigger in some cases - not because of lack of understanding, but of lack of confidence in what the *right* answer was *supposed* to be (e.g., MI16T, MI5T, MI10F, V5A18).

In some cases, he clearly held one of the tested misconceptions and the instrument was able to capture that (e.g., Misinterpretation #1 was answered incorrectly in both presentations: FALSE and Vignette formats; Misinterpretation #2 was answered incorrectly in two of three presentations: FALSE and Vignette formats; Misinterpretation #5 was answered incorrectly in two of three presentations: FALSE and TRUE formats).   In other cases, when he demonstrated a correct conception, he seemed to perform better with P/CP and RV because the two options allowed him to account for things like *not necessarily* that were causing him conflict in the T/F section.

The final participant profiled was Shelley. She was judged as having high understanding of quantitative methods in general and *p*-values in particular. Presented here are a few sample transcript excerpts:

- To me it seems like you have more people now and your wait time got closer to 30 minutes, they are both about 1%, but I would say that they are not necessarily in agreement because you are trending the other way and if you get more participants you might actually see that the wait time is a result of your sample since you only had so many customers. I would say you should NOT assume these results to be in agreement.
- *At least as small*…you mean, they will produce a *p*-value that is the same or smaller? I agree with the CP because of the *at least* as small….you would hope to get similar results but I don't know if you can say *at least* as small…[after discussion]…I think just because you have 2 studies testing the same hypothesis does not mean that you are going to get similar *p*-values or approximate or smaller than - particularly smaller than I find weird...so if you ask *approximate*, it might be a better way to say it.
- …so, *unusual* means…what? Let me read the other one. Are we talking about the difference between practical significance and statistical significance? You can get a significant difference in pre/post scores but the student only improved by 2 points. The word *unusual* is a bit off.   In Point, you said what you meant. In CP, it took time to parse out the difference in the statements. Maybe use *clinical interest*.

She is able to discuss ideas fluently and make commentary on mistakes she has observed others make. Her decision-making was calculated and confident. It was easy for her to imagine/devise alternate scenarios than those described and was not stymied upon encountering statements that needed qualification (e.g. *not necessarily*). Her ease with the subject matter permitted her to both engage with the instrument as a respondent and evaluate the instrument on its technical merit simultaneously, even going so far as to offer specific suggestions as to how individual items might be improved (e.g., MI11T, MI17F, MI5F, MI7p, MI9p, MI18cp, V8A6, V8B13, V8B7).

Her strong knowledge of the content allowed her to (properly) infer the likely meaning of key words and phrases that seemed to baffle other participants. In her case, the instrument did not obfuscate her ability to demonstrate knowledge. I think, on the contrary, her independent decoding actually served to reveal her depth of understanding (e.g., V2B17, MI18cp). If substantial scaffolding is necessary for engagement with the items, perhaps it is not necessarily an indication of poorly constructed items but rather that these topics are on the periphery of (or beyond) comprehension.

Apart from identifying flaws with the instrument design and the items themselves, the think-aloud interview setting afforded the opportunity to estimate the depth of *p*-value understanding of the participants. In the absence of a true convergent validity measure, the level of methodological understanding and *p*-value fluency exhibited in discussion with the researcher supplied evidence that was used in ranking the participants based on the researcher's (subjective) assessment of each respondent's relative standing within the group. These rankings were then compared to performance on the instrument – both in total and by subscore – as a means of gauging the consistency of performance on the assessment with fluency.

Inarguably, Shelley was the most knowledgeable of the participants and Dana was the least. The ordering of the other three was a bit more difficult to parse. Upon initial inspection, there was not much to separate them. All seemed to have similar profiles: a combination of a few firmly entrenched misconceptions coupled with a strong understanding of most of the concepts. All had gaps in their knowledge but were capable of identifying these weaknesses and most ideas were in their zones of proximal development as their total statistical understanding allowed them to meaningfully engage in conversations with the researcher when clarification was provided. Unlike with Dana, for these participants, most of the misinterpretations seemed to be a result of an emergent understanding of the concept that could be readily ameliorated with a bit of clarification.

Ultimately, the ranking decision was determined based somewhat on perception of experience with quantitative research. Mike had the upper hand in this criterion as his commentary was often peppered with remarks about his identification of the misconceptions of others. His ability to recognize and discuss in depth methodological flaws and statistical foibles suggested that peer review was something with which he was quite familiar. Not insinuating that this capability renders him immune from mistakes of his own, but his experience with this aspect of peer review would likely give him an advantage in the type of assessment under investigation.

Abigail and Babs were the hardest to separate, but in the end, the edge was given to Babs based on a slight (perceived) advantage in experience. Ironically, Babs actually never supplied a proper *p*-value definition (something Abigail very nearly did), but somehow that did not cause her to suffer from as many misconceptions as one might expect. Her commentary suggested a stronger familiarity with quantitative methods in general, relative to Abigail, that allowed her to better understand the research scenarios presented in the vignettes and afforded her the capability of devising more realistic contextual examples for herself in the cases when the misconceptions were provided without context in the assessment (i.e., P/CP, T/F).

The final rankings can be seen in Table 16 along with the participant scores in total and by subsection. For the most part, the scores were consistent with expectation based on rank. Within the P/CP and RV sections, the scores were in perfect alignment. The T/F section showed an interesting inconsistency in that Mike scored lower than both Babs and Abigail. Rather than serving as evidence that perhaps the rankings were incorrect, this more than likely suggests a flaw with the T/F items. As previously identified in his profile, Mike was preoccupied with 'gaming the assessment' and trying to locate the 'trick' word in the statement that made it false. This preoccupation with the 'right' answer interfered with his ability to display his true understanding revealed in discussion. The inconsistency in the T/F scores contributed to a slight inconsistency in the Total score (Babs outscored Mike), but otherwise, this analysis can be seen as supporting the validity of the instrument.

Table 4.12 – Participant Ranking Cross-Tabulation

Name	Level of Knowledge Ranking	T/F Score	P/CP Score	RV score	Total Score
Shelley	1	0.9375	1.0000	0.9444	0.9512
Mike	2	0.6875	0.8571	0.8333	0.7805
Babs	3	0.7500	0.8571	0.8333	0.8049
Abigail	4	0.7500	0.5714	0.6667	0.6829
Dana	5	0.6250	0.5714	0.6111	0.6098

4.1.2.9 Phase IIB Usability Data

The Phase IIB participants were asked to complete the same exit interview items as the IIA participants and the results were mostly consistent with those previously obtained. The majority of participants indicated no issues with accessibility (3/4) or with the survey navigation (3/4) or font (4/4)[20]; however, only 2/4 participants were satisfied with the number of items per

[20] As mentioned previously, 4 of the 5 participants took the assessment electronically; 3 on a laptop and 1 on a cell phone. The 5th participant was unable to use an electronic device due to technical issues (unrelated to the instrument itself) and took a paper-based version. Her responses for these items were not recorded as it was not applicable.

page (the remaining 2/4 indicated there were “too many”). Participants generally felt that the directions were easily understood (5/5 agreed for the T/F section, 4/5 agreed for the P/CP section, and 5/5 agreed for the Vignette section). There were no identified issues with the reading level of the instrument (5/5 agreement) nor the length of vignettes or ratio of items per vignette (5/5 agreement for both). The primary source of contention was in the ordering of the sections within the instrument. Only 2/5 participants agreed with the current formatting (T/F – P/CP – RV); the remaining 3 responses were all distinct (RV – P/CP – T/F, T/F – RV – P/CP, RV – T/F – P/CP). Due to the nature of the think-aloud interview format, average response time was inestimable. In terms of incentivization, the results were again inconclusive as to the best choice. While all five participants were willing to consider alternative inducements to payment, not all options were deemed equally viable. Homework and Extra Credit received unanimous approval (5/5 for each), but there was less support for the raffle option as only 3 of 5 participants were willing to accept this incentive.

4.1.2.10 Implications of the Pilot Test Results

In terms of the instrument development process, the field test version of the instrument was informed by the pilot test results in three main areas: recruitment, usability, and performance. Considering Phase IIA first, the surprising finding here was that the graduate listserv was markedly effective at recruiting participants, both in terms of sample size and variety of demographics. In terms of usability, the primary takeaway was that participants were not dissuaded by the response time (an initial researcher concern, especially in the absence of paid participation), and furthermore, this variable had the potential to improve based on suggested item edits. Regarding incentivization, there was no clear answer, indicating that any of the choices would be feasible going forward, but a slight edge was shown for the Homework option.

Based on the performance data, the primary concern was that the True/False section might not be functioning appropriately due to low correlations with other subscores, total score, and convergent measures indicating that major modifications would be necessary for retention of this subsection – either in individual item wording or in formatting in general. Lastly, while not necessarily imperative, the criticism of redundancy could be addressed by removal of one of the vignette sections. Reliability analysis indicated that the Odds subset was outperforming the Evens (evidence presented in Table 4.9 and Appendix M) and thus removal of the Evens subset might be considered.

Considering Phase IIB, the principal contribution of this phase was in the wording of individual items. The identification of tricky words and phrases informed the next iteration of the items. This is addressed in detail later in this chapter. Regarding recruitment, the techniques applied here are not feasible in the field test and thus no practical implications can be derived. In terms of usability, the cognitive labs revealed the shortcomings of the Qualtrics instrument on a cell phone plaT/Form and suggested that future participants be urged to complete the assessment on a laptop for ease of readability and navigation. As for incentivization, the Phase IIB participants basically echoed the sentiments of their IIA counterparts indicating that all the inducements were essentially equivalent. The performance data, while limited in scope due to sample size and researcher interference, similarly indicated that the True/False section should be reconsidered.

4.1.2.11 Instrument Modification Summary and Revised Test Blueprint

In accordance with the feedback received from the Pilot Test, a list of proposed modifications was presented to the dissertation committee for their consideration. The modifications were categorized as being General (relating to the structure of the instrument) or Specific (relating to the wording of individual items).

The first of three major modifications proposed was the suggestion to remove the T/F section from the instrument and reword all items in a P/CP format. The justification for this was based on a variety of factors. Of primary concern was the poor performance of this section (as already discussed) in both phases of the pilot test. Owing in part to the difficulties outlined in the presentation of the items design, negation of the statements was not entirely straightforward and ultimately additional items (not contained in the theoretical framework) were required in order to balance the instrument. Removal of this section would help to alleviate these concerns.

As discussed with the item panel, one of the original ideas for the instrument was to present the items entirely in P/CP format (no traditional T/F) as a means of parallelism between the with/out context items (the RV items were designed to be P/CP from the outset); however, there was concern that the instructions might be confusing for this type of task and stymie and/or frustrate respondents. The pilot test dispelled this notion as the majority of participants indicated no difficulty with the task expectations and some even indicated a preference for this format as it allowed them to abandon the search for key words and phrases that might be making the statement false and focus on determining which of the two statements seemed the more appropriate interpretation. Another advantage to this suggested modification is the internal consistency of the instrument as all sections would now share the same task instructions. This modification was implemented.

The second of the three major general modifications proposed was the suggestion to alter the length of the instrument by modifying the presentation of the research vignettes. The justification for this was that it would reduce response time (in the hope of receiving more participants and also more completed responses) and would eliminate redundancy (a common complaint from the pilot test). The PV1 version of the instrument currently contained four vignette pairs split into even and odd subsets (2A/B & 8A/B; 3A/B & 5A/B respectively). The Phase IIA participants completed all four pairs whereas the Phase IIB participants only completed either the odd or even subset (randomly selected). Since each odd/even subset independently captured all 18 misinterpretations, the internal redundancy was by design with an eye to test reduction in future iterations when there would be sufficient data to reconcile the tradeoff and negotiate the balance between test length (short enough to please participants) and depth of diagnostic information (long enough to diagnose specific misconceptions).

Three proposals were considered: the PV2 iteration of the instrument would contain only one subset (odds or evens), the PV2 iteration would contain one complete pair (both odds or both evens) and one cross pair (one of the opposite parity) for a total of 3 vignette pairs tested, the PV2 iteration would have parallel forms in which one contained only the odd subset and the other contained only the even subset and participants would be randomly selected to receive a particular version of the assessment. The advantage to the first option is that it provided the most data per item with the fewest number of items. The disadvantage is that it limited the conclusions for individual ability estimates with the removal of replicates. The advantage to the second option is that the extra redundancy offered some replicate measures but only for a subset of the 18 misinterpretations and any marginal value contributed by these additional items would likely not offset the reduction in response rate and number of completed responses incurred by the additional test length. The advantage to the third option is that allows both sets of vignette pairs to remain under consideration in the item pool to be used (in part or parcel) for future iterations of the instrument while simultaneously supplying increased evidence for the validation plan. The disadvantage to the third option is that the split option reduces the sample size by half possibly affecting the model parameter estimates and limits the generalizations therein. This modification was implemented by way of removal of the even subset based on the justification that the odd subset had performed superiorly in the pilot test.

The third of the three major general modifications proposed was the notion of changing the items from dichotomous to polytomous. An important phenomenon that was revealed during the cognitive labs was that participants had many idiosyncratic methods for selecting answers in the absence of a clear preference. Without the interview transcript (in an online session), there is no way to filter out the guesses from the genuine responses. The proposed wording for the polytomous items included three choices: TRUE = I agree with this idea, FALSE = I disagree with this idea, I DON'T KNOW = I have never heard of this idea or I don't understand this idea.

The justification for this modification was that sometimes participants do not have enough knowledge of the subject to understand the item or to differentiate between the alternatives; a third option can reveal that position. This third option could filter the differences between what people incorrectly believe and what they do not know. From a scoring standpoint, incorrect guesses are not really that different from purposefully selected incorrect responses corresponding to misconceptions (i.e., wrong is wrong); however, from an interpretive standpoint, there is an important distinction. Researchers who firmly hold a misconception are unlikely to consider their viewpoint as incorrect and are thus likely to perpetuate this thinking in their own research and in peer review; whereas, an idea that is foreign to them or confusing is less likely to be propagated as it is more likely to be avoided. In terms of the second research question, *What do the results of the KPVMI-1 administration tell us about the current level of p-value fluency among doctoral students nationally?*, it is an important distinction between ignorance and misconception/misinterpretation. The former can be theoretically ameliorated with further training, the latter not necessarily so. This modification was not implemented primarily because there was concern that the third choice might become a metaphorical *dumping ground* where participants would default when not motivated to engage with concepts eluding their working familiarity and that response sets might be sparse of a reasonable number of actual responses.

In addition to the three major general modifications considered, there were some minor general modifications that were implemented. Firstly, with the removal of the T/F section, the most preferable arrangement (based on user feedback) of the remaining two sections was P/CP first followed by the research vignettes. Next, the formatting of particular items were updated to emphasize the differences in the paired statements based on user feedback from the pilot study (i.e., Phases IIA and IIB) indicating that some item pairs required multiple read-throughs in order to determine the nature of the distinction. After considering colored text, bold font, and underlining, the committee settled on a combination of capitalization (of key words like NOT) and underlining key phrases. The final minor modification affected the item presentation. During the pilot testing, individual randomization (achieved in Qualtrics) was utilized. The committee felt this was an unnecessary precaution as cheating was unlikely given the nature of the recruitment and thus this was eliminated. For each P/CP pair, a coin flip determined whether the correct or incorrect statement would be listed as *Point* and then a random number generator was used to determine the order in which the misinterpretations were presented to the user. This was after dismissing the original notion of presenting the misinterpretations in numerical order (i.e., #1-#18) based upon inspection of the response pattern (i.e., too many P or CP consecutively could alarm respondents and cause them to modify their responses accordingly).

Once general modifications had been implemented, specific modifications to individual items were performed. Primarily this involved the clarification or elimination of specific phrases identified by participants as causing confusion or consternation. This included, but was not limited to, *perfect agreement*, *observed association*, and *totality of evidence*. Lastly, the following maintenance items were performed: fixed all typos identified in the pilot test, changed percent to decimals where appropriate for consistency, similarly reworded where appropriate to match text with inequality symbols, and eliminated the redundant presentation of hypothetical scenarios in two-sample items.

The objective of Phase II of this research was to generate the next iteration of the *KPVMI* instrument. Starting with a preliminary version based on the item panel results, the internal structure, item stems, and research vignettes were modified in accordance with the pilot study feedback and were consolidated into an updated version (PV2) suitable for field testing with actual subjects from the target population. The updated test blueprint is presented here (Table 17) and a complete log of all item modifications and comments therein can be found in Appendix O.

Table 4.13 – Test Blueprint Revisions

Modifications to Test Blueprint (PV1 → PV2)					
	Process Level				
Item Content	**Context-Free**			**In Context**	
	~~TRUE~~	~~FALSE~~	P/CP	Odd Vig	~~Even Vig~~
Misinterpretations of single P-values	~~7~~	~~6~~	**14**	14	~~14~~
MI1		~~1~~	**1**	1	~~1~~
MI2		~~1~~	**1**	1	~~1~~
MI3			1	1	~~1~~
MI4		~~1~~	**1**	1	~~1~~
MI5	~~1~~	~~1~~	**1**	1	~~1~~
MI6		~~1~~	**1**	1	~~1~~
MI7			1	1	~~1~~
MI8	~~1~~		1	1	~~1~~
MI9			1	1	~~1~~
MI10	~~1~~	~~1~~	**1**	1	~~1~~
MI11	~~1~~		**1**	1	~~1~~
MI12	~~1~~		**1**	1	~~1~~
MI13	~~1~~		**1**	1	~~1~~
MI14	~~1~~		**1**	1	~~1~~
Misinterpretations of P-value comparisons and predictions	~~2~~	~~1~~	**4**	4	~~4~~
MI15			1	1	~~1~~
MI16	~~1~~		1	1	~~1~~
MI17	~~1~~	~~1~~	**1**	1	~~1~~
MI18			1	1	~~1~~

Test Blueprint – PV2 Version	
Process Level	
Context Free	**In Context**
P/CP	Odd Vig
14	14
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
4	4
1	1
1	1
1	1
1	1

Note that this table shows the new blueprint (PV2) as reflected in changes to the old version (PV1). Items added are marked in* **red *and items removed have been ~~stricken~~. The current test blueprint shows a total of 36 items - 18 each with/out context.*

4.1.3 Phase III of Instrument Development (Field Test)

Phase III of the study gathered empirical evidence in support of the measurement model and validation plan through field testing. This portion of the research was conducted in two stages: first using students enrolled in a large interdisciplinary graduate research methods course at Virginia Tech (n = 79), and later in a national sample (n = 207). Data from both samples were used individually and collectively as appropriate to the validation efforts. The conclusion of this phase of the research culminated with the identification of a subset of items to serve as the (current) final[21] version of the instrument (PV3).

4.1.3.1 Phase IIIA Overview

Phase IIIA was the local field test. Designed to collect baseline data for item analysis, the target sample size was 100 participants recruited from a VT methods course. A copy of the Qualtrics version of the instrument as presented to participants can be seen in Appendix P.

4.1.3.2 Phase IIIA Participants

The class roster of the Spring 2018 (VT) section of STAT5616 was the basis for the sampling frame. Justification for this recruitment decision (as explained previously in Section 3.2.4.1) included the following considerations: (1) the large class size suggested a reasonable response rate would yield an adequate number of participants, (2) the course is the capstone of a year-long methods sequence and thus the students should possess the requisite knowledge to engage meaningfully with the instrument, and (3) the course is a service course offered to graduate students across departments and would theoretically yield a diverse representation of programs of study. The target sample size was n = 100 and recruitment efforts yielded 79

[21] The *final* version of the instrument in terms of *this* research study. As discussed in Chapter 5, the instrument validation is a work in process and future data collection and instrument revision is planned.

completed[22] responses spanning 37 programs of study. Additional demographic information regarding status, gender, and race can be seen in Table 4.14. Of particular interest here is the noticeable right skew in the status variable, i.e., there are very few senior PhD students and an unexpectedly large contingent of Master's students.

Table 4.14 – Local Field Test Participant Data

Status	Master's	1st yr PhD	2nd yr PhD	3rd yr PhD	4th yr PhD	5th yr PhD
Frequency	44	16	12	4	2	1
Percent	55.7%	20.3%	15.2%	5.1%	2.5%	1.3%
Sex/Gender	**Male**	**Female**	**Not Provided**			
Frequency	47	30	2			
Percent	59.50%	38%	2.50%			
Race/Ethnicity	**Caucasian/ White**	**Asian**	**Black/ African**	**Hispanic/ Latino**	**Other/ Mixed**	**Not Provided**
Frequency	36	21	2	6	7	7
Percent	45.60%	26.60%	2.50%	7.60%	8.90%	8.90%

4.1.3.3 Phase IIIA Usability Data

Relative to previous iterations of the instrument (i.e., in the pilot study), the field test version had a much abbreviated exit interview portion; thus, the usability data was limited. The primary consideration here was response rate and response time. The response rate was remarkably high with 98% of the enrolled students participating (99/101) and approximately 80% of those responses complete. (Of the 20 incomplete responses registered, 15 had no items data – only consent and then survey was exited). The remaining four partial responses (60-85% completion) were removed from consideration on the grounds that the paucity alone relative to the sample size gave no compelling reason for inclusion. Response time was reasonable with a

[22] Completed here implies of all *assessment* items; exit interview items were not required to be answered for a response to be considered.

mean of 44 minutes and a median of 26 minutes – a slight improvement over the pilot test version. User feedback was solicited via textbox; however, the responses were not analyzed at this stage of the research and will be discussed in Chapter 5. The complete list of usability comments can be seen in Appendix Q.

4.1.3.4 Phase IIIA Items Data

As was the case for the pilot test, the purpose of this phase of the field test was not to analyze the performance of the respondents on the items, but to assess the performance of the items themselves. That being said, a cursory glance at respondent performance is warranted to verify that scores are not inappropriately high or low relative to expectation. In this instance, overall respondent performance on the instrument suggested moderate levels of *p*-value fluency with total scores indicating respondents correctly answered slightly better than half of the items, on average – scores that are a bit poorer than those of the field test participants, but perhaps not unsurprising considering the sample (i.e., enrolled students seeking extra credit vs. researcher-selected participants). Specific scoring details, in total and by sub-section, can be seen in Table 4.15 and in Figure 4.11

Table 4.15 – Phase IIIA Scoring Statistics

Descriptive Statistics					
Instrument Section	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
P/CP Score	79	0.28	0.83	0.5211	0.13704
RV Score	79	0.17	0.89	0.5267	0.15860
Total Score	79	0.28	0.83	0.5239	0.12805

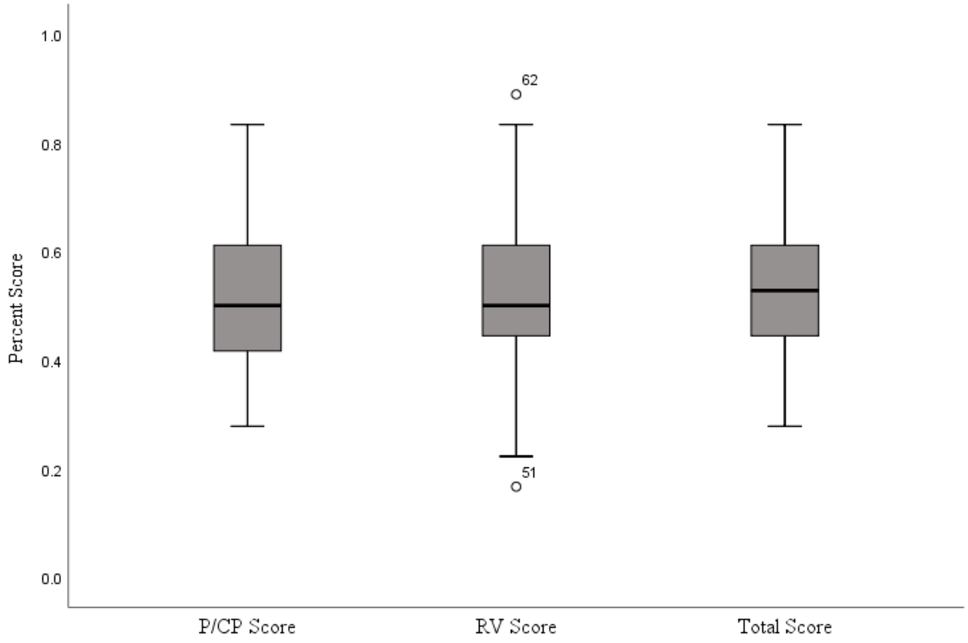


Figure 4.11. Phase IIIA Distribution of Scores by Section

Turning our attention to the *item performance*, this will be addressed on the whole and via subsection. Considering the P/CP section first, Figure 4.12 shows the percentage of correct answers by item arranged in numerical order. Items that received more than 60 correct responses (n = 2) or less than 20 correct responses (n = 1) were flagged for consideration, but were not eliminated solely on this criterion at this stage of the analysis.

A similar examination of the research vignette items is displayed in Figure 4.13, arranged in numerical order. There were no items that were flagged for being too *hard* (i.e., < than 20 correct responses), but there were a couple (n = 2) flagged for potentially being too *easy* (i.e., > 60 correct responses).

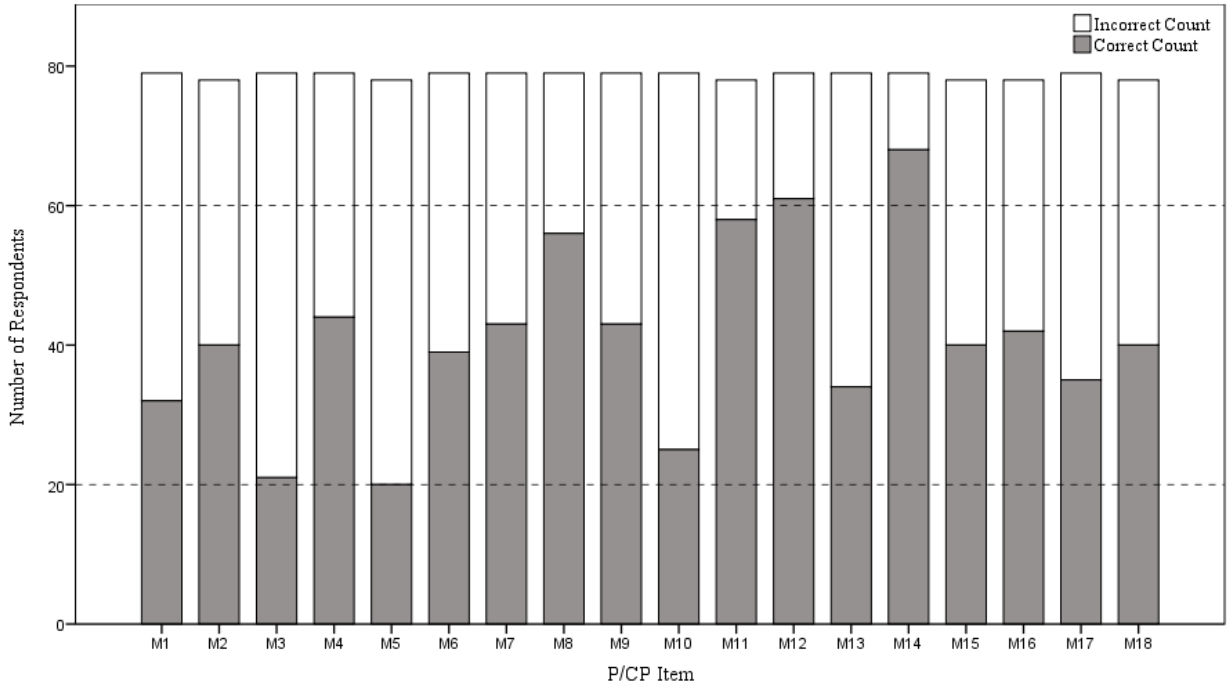


Figure 4.12. Participant Performance – P/CP Section

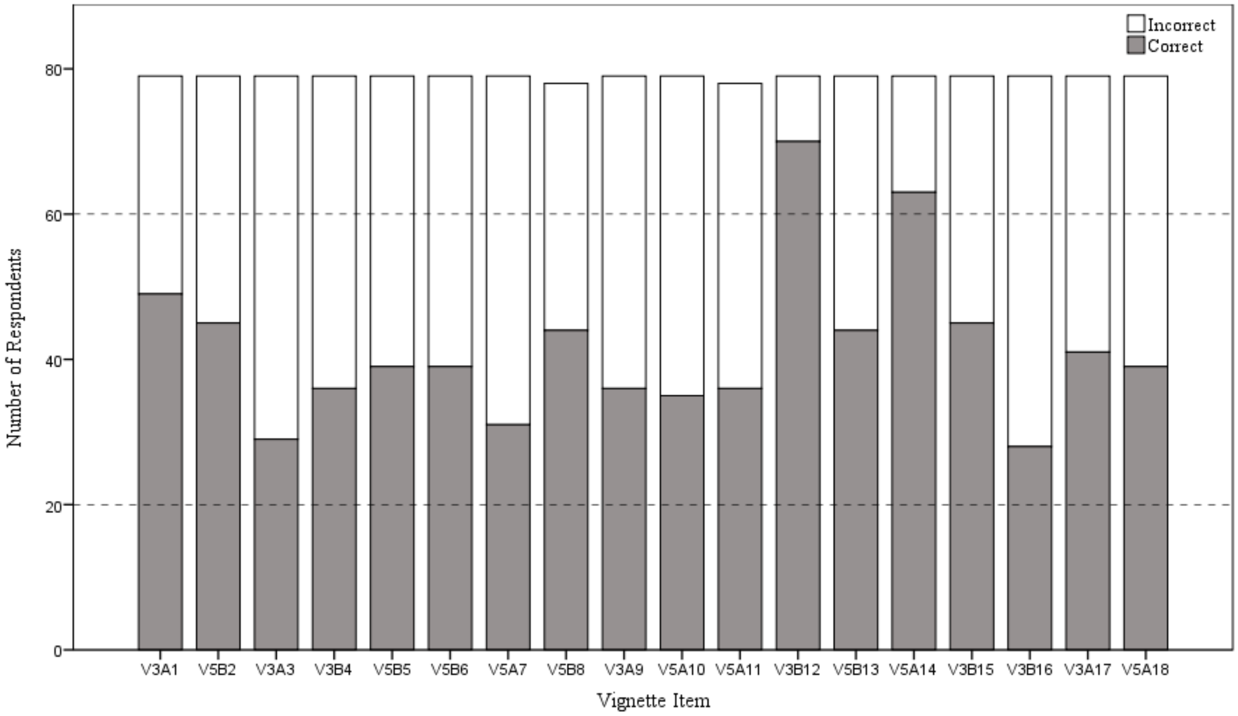


Figure 4.13. Participant Performance – Vignette Section

As with Phase II data, the internal consistency of the items across formats was investigated. Pearson correlations were computed to measure the strength of the relationship of each subsection score with the total score. In this assessment, high positive correlations (i.e., close to 1.0) were desired as this would indicate that higher scores on a given subsection correlates with higher scores overall. Table 4.16 shows the correlation matrix (cells display *r* with the *p*-value in parentheses). The research vignettes showed the stronger correlation with total performance as compared with that of the P/CP section; however, the difference was small and likely inconsequential. The strong positive correlation of both sections with the total score, and ordering therein, is consistent with those results seen in Phase IIA.

Table 4.16 – Correlation Matrix – Phase IIIA

Correlations

	Total Score	P/CP Score	RV Score	Experience
Total Score				
P/CP Score	.844 (< .001)			
RV Score	.886 (< .001)	.498 (< .001)		
Experience	.215 (.057)	.205 (.070)	.170 (.134)	
Status	.271 (.016)	.182 (.109)	.280 (.013)	.275 (.014)

Also in keeping with the Phase II analysis, an investigation of the correlation between scores and coursework as a proxy for convergent validity was performed. Theoretically, students who have completed more relevant coursework should have a better understanding of the statistical ideas and suffer from fewer misconceptions and misinterpretations (thus high, positive correlations would be desired for this measure). Neither the total score (r = .215, p = .057), nor either subsection (r = .205, p =.07; r = .170, p = .134), correlated satisfactorily with this measure. These unexpectedly low correlations are inconsistent with results seen in Phase II and a cause for concern. That the correlations are positive (albeit low) is somewhat reassuring on face because it implies that the ‘more coursework = higher scores’ relationship is maintained; however, the marginal contribution of additional coursework implied by this analysis suggests that either fluency does not improve with additional training or that coursework is an inappropriate proxy for a convergent validity measure.

Linear regression was employed to model the relationship between courses completed and total score, merely as a means of comparison with the Phase IIA results (not based on the expectation that this would be statistically significant or meaningful in and of itself). The suggested relationship ($\beta_0 = .465, SE_{\beta_0} = .034$; $\beta_1 = .029, SE_{\beta_1} = .015$; ANOVA F = 3.739; p = .057) indicated that a person with no experience would score about 47% on average with an expected increase of about 3% for each additional course completed – nearly 1/3 of the per course contribution suggested in the pilot test . See Figure 4.14 for details.

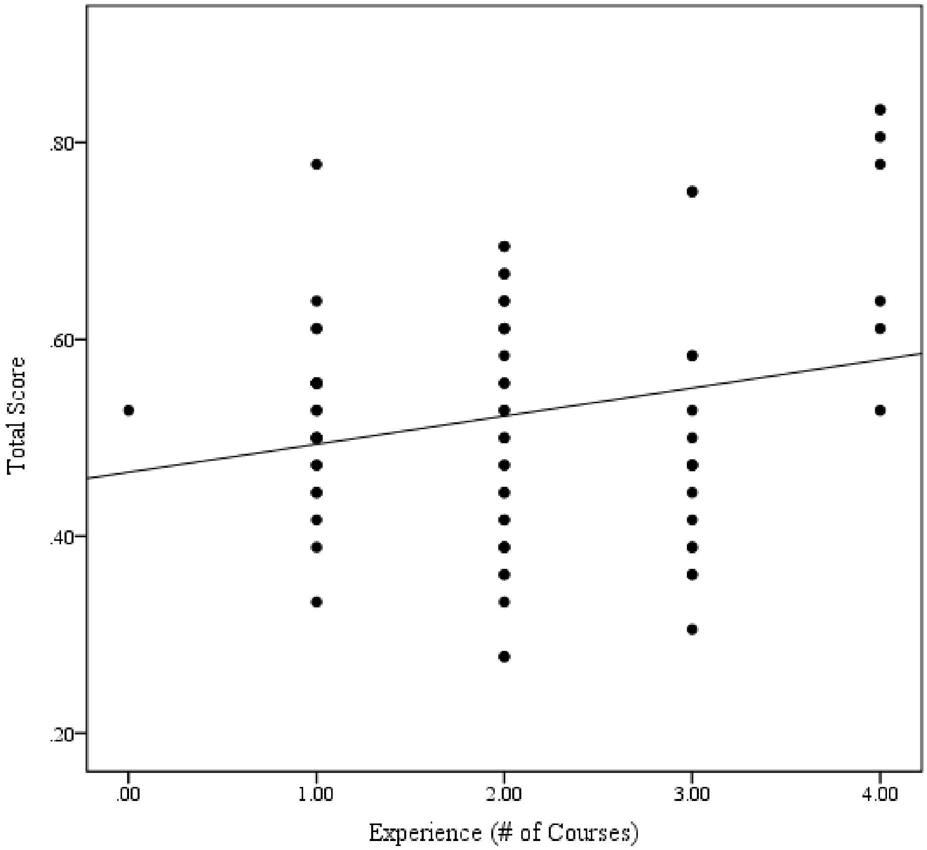


Figure 4.14. Regression Results – Phase IIIA

As with Phase IIA, the final assessment of instrument performance was a check on internal reliability, both in total and by subsection. When using Cronbach's alpha, the closer a value is to 1.0, the more reliable the measure. Conventional wisdom suggests that a value of 0.70 is the lowest acceptable threshold (Nunally, 1978). As can be seen in Table 4.17, this threshold has not been met for the instrument as a whole or by subsection for either the P/CP or Odd vignettes.

In diagnosing the reliability of a collection of items, an important follow-up analysis is to examine the *Corrected Item-Total Correlation* and/or *Cronbach's Alpha if Item Deleted* measures. These diagnostics will indicate which items, if any, are problematic and adversely affecting the instrument's reliability. In the case of *Cronbach's Alpha if Item Deleted*, values that exceed the alpha of the collection of items suggest that the instrument would be improved if these items were removed. In the case of *Corrected Item-Total Correlations*, negative values indicate that there is an inverse relationship between performance on this item and the total score. In this analysis, items were classified as 'poorly functioning' if the *Corrected Item-Total Correlations* fell below the 0.05 threshold. This criterion identified six potential issues with the P/CP items and three with the vignettes (details can be found in Appendix R).

Table 4.17 – Cronbach's Alpha Data Phase IIIA

Subset	# of Items	Alpha	Diagnosis	Items
P/CP	18	0.328	6 items are poorly functioning	M12, M13, M18, M14, M11, M2
RV odd	18	0.511	3 items are poorly functioning	V3A9, V3B12, V5A14
All items	36	0.599		

Promising initial results after the pilot test gave no reason to suspect that the field test, utilizing an improved version of the instrument, would not fare as well. While there were aspects that remained consistent with Phase IIA, there were enough indicators that were demonstrably worse to warrant further investigation. The principal difference between the samples, apart from the obvious size discrepancy, was the qualifications of the participants. As mentioned in section 4.1.3.2, an unexpectedly large contingent of Master's students participated in Phase IIIA. Target demographic notwithstanding (i.e., PhD students nearing the completion of their training), there was initially no reason to suspect that the inclusion of Master's students in the data pool would be inimical to the aim of assessing item and instrument performance, especially considering there were Master's students amongst the Phase IIA participants. Remember, inferior *student* performance is not only permissible but expected and desirable (less experience and coursework should translate to less fluency) from an instrument validation standpoint; it was the inferior *item* performance that was disconcerting and unanticipated.

Owing to the recognition that 44 of the 79 participants were Master's students, a secondary investigation was initiated as a means of accounting for the inconsistencies in item and instrument performance in which each of the aforementioned analyses was repeated on a split data set (Master's vs. Not). The results are presented in their entirety in Appendix S and are discussed here in brief.

In terms of student performance, on the whole and by item, discrepancies were either non-existent or inconsequential; however, striking differences in item performance exposed between the sub-populations were sufficient to justify this investigation. The correlation between course completion and performance score was suspiciously low for the composite dataset and disaggregation revealed that an inappropriate relationship (r = -.235, p = .125) in the Master's subset was the culprit (PhD subset was a respectable r = .508, p = .002). The linear regression models confirmed this result with a (insignificant, p = .125) predicted *decrease* of approximately 3% for every additional course completed in the Master's subset; a reasonable predicted *increase* of approximately 7% was recorded for the PhD subset. Reliability analysis showed no deviation from this trend. The PhD subset showed far superior item performance across the board. Details are shown in Table 4.18. In light of these findings, and in accordance with the nature of the research questions, there was no compelling justification for expansion of the target demographic to include non-PhD students either in future iterations of data collection or in subsequent analysis of data in hand. Thus, all Master's students in the Phase IIIA dataset were removed for cause and excluded from consideration in future data collection.

Table 4.18 – Cronbach's Alpha Data Phase IIIA by Population Subgroup

Master's Subset				
Subset	**# of Items**	**Alpha**	**Diagnosis**	**Items**
P/CP	18	0.253	9 items are poorly functioning	M1, M13, M18, M9, M16, M14, M11, M7, M17
RV odd	18	0.248	9 items are poorly functioning	V3A9, V3A17, V3B12, V5A10, V5A14, V5A18, V5B6, V5B8, V5B13
All items	36	0.454		
PhD Subset				
Subset	**# of Items**	**Alpha**	**Diagnosis**	**Items**
P/CP	18	0.424	5 items are poorly functioning	M12, M13, M18, M14, M2
RV odd	18	0.665	3 items are poorly functioning	V3A9, V3B12, V5A14
All items	36	0.71		

The original intention of this research was to use the local field test (Phase IIIA) for instrument validation and to use the national sample (Phase IIIB) in pursuit of the second research question. A smaller than expected sample size (n = 79) for the VT sample rendered this approach unlikely to yield stable parameter estimates and satisfactory item analysis diagnostics. Further reduction of the dataset (n = 35) by elimination of the Master's students made this aim all but impossible; thus, it was decided to postpone certain aspects of the validation plan until both sets of data (i.e., Phases IIIA & IIIB) were in hand and to pool them when appropriate and/or necessary. As such, certain components of the validation plan (e.g. external validity) were not undertaken until after Phase IIIB, while other diagnostics (e.g. classical item analysis) were performed on Phase IIIA data and then repeated after Phase IIIB. Accordingly, further discussion of instrument validation procedures will be discontinued at this time and resumed in Section 4.1.4.

4.1.3.5 Phase IIIB Overview

Phase IIIB was the national field test. Designed to collect additional evidence in support of instrument validation and to provide baseline data in pursuit of the second research question, the target sample size was 400 participants recruited from doctoral programs at R1 institutions (as classified by the Carnegie Classification of Institutions of Higher Education) nationwide. In response to the small sample size, and unexpected reduction therein, no modifications to the instrument structure or item revisions took place between Phase IIIA and Phase IIIB. With an eye to merging and/or comparing/contrasting the field test results by phase, it was necessary that all participants test under identical conditions; thus, the national sample participants completed the same version of the instrument (PV2) as used in the local field test (Appendix O).

4.1.3.6 Phase IIIB Participants

The complete list of 115 R1 institutions nationwide was the basis for the sampling frame (shown earlier in this document, Table 3.3). Participants were recruited indirectly via local dissemination efforts coordinated by graduate student organizations. The target sample size was n = 400 and recruitment efforts yielded (a disappointing) 207 responses spanning at least 16 unique institutions[23]. Participating institutions included: Brown University (7), Case Western Reserve University (7), Northwestern University (4), Penn State University (5), Purdue

[23] Only 87 respondents identified their institutional affiliation. For the remaining 120 respondents, there was no way to recover this information and thus no way to state with accuracy exactly how many institutions actually participated. There were no minimum thresholds for an institution's participation; any and all valid responses from any R1 institution were considered.

University (6), Rice University (4), Tulane University (8), University of Arizona (10), University of Connecticut (10), University of California at Berkeley (2), University of California at Santa Barbara (9), University of California at Riverside (10), University of Louisville (3), University of Buffalo (1), and Washington State University (1). Additional demographic information regarding Status, Sex/Gender, and Race/Ethnicity can be seen in Table 4.19. The sample is primarily white and female; however, unlike with the local field test data, there is a nice distribution across levels (i.e., year in school). Of particular interest with this data is the unexpectedly large proportion of participants who declined to provide any demographic information (119 missing Status, 119 missing Sex/Gender, 121 missing Race/Ethnicity).

Table 4.19 – National Field Test Participant Data

Status	Master's	1st yr PhD	2nd yr PhD	3rd yr PhD	4th yr PhD	5th yr PhD
Frequency	7	19	13	13	13	17
Percent	3.4%	9.2%	6.3%	6.3%	6.3%	8.2%
Sex/Gender	**Male**	**Female**	**Not Provided**			
Frequency	28	60	119			
Percent	13.5%	29.0%	57.5%			
Race/Ethnicity	**Caucasian/ White**	**Asian**	**Black/ African**	**Hispanic/ Latino**	**Other/ Mixed**	**Not Provided**
Frequency	60	15	1	5	5	121
Percent	29%	7.2%	0.5%	2.4%	2.4%	58.5%

4.1.3.7 Phase IIIB Usability Data

Due to the abbreviated exit interview portion (as explained for Phase IIIA), the usability data was limited. The primary consideration here was response rate and response time, both of which were disappointingly low (see Figure 4.15). Nearly half of the institutions did not respond to the solicitation and only 23% of those who did respond indicated a willingness and intention to participate. Common reasons offered for rejection included IRB concerns and lack of an appropriate dissemination method and/or precedent for such a request.

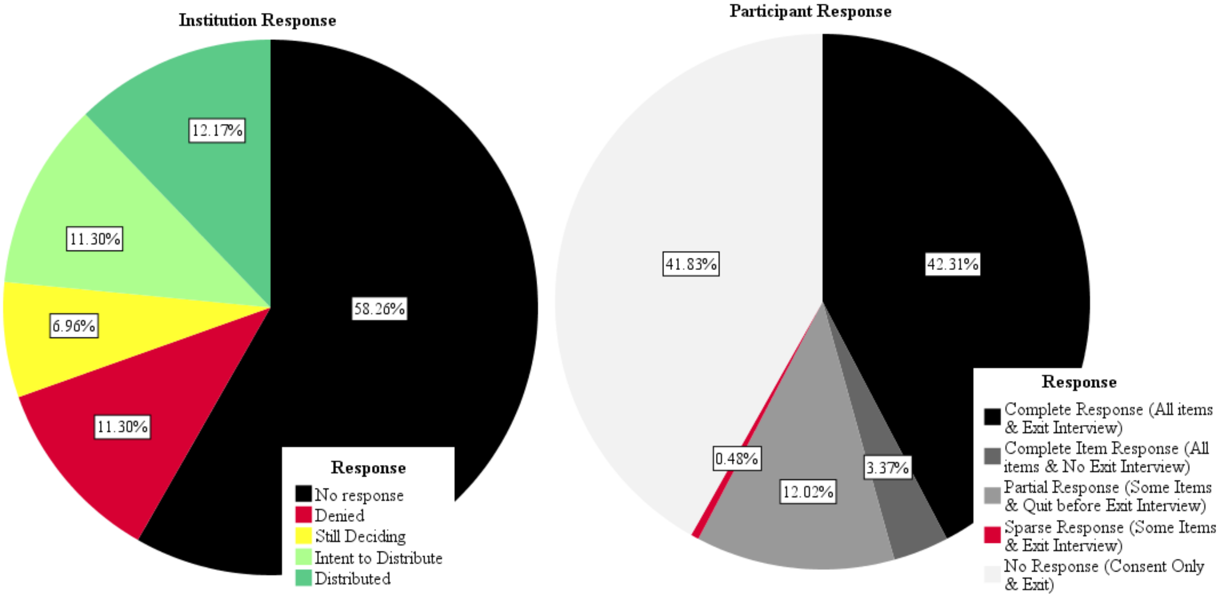


Figure 4.15. Phase IIIA Response Data

Within the participating institutions, a reliable measure of response rate was incalculable without knowing the size of the graduate student population or reach of the dissemination efforts. As an alternative measure, the response rate as a function of *completed* responses was calculated. Results were essentially binary here: nearly as many *Non Responses* were recorded as *Complete Responses* (41.83% vs. 42.31%); sparse and truncated responses were conspicuously rare. These numbers were substantially weaker than in Phase IIA, but that is relatively unsurprising considering the nature of the incentivization (*potential* raffle prize vs. *guaranteed* extra credit). The descriptive statistics for response time are not meaningful due to the high proportion of incomplete responses (mean = 75.32 minutes, median = 12.88 minutes). Recalculations performed on a trimmed data set (complete item responses only and outliers[24] removed) yielded

[24] There were 3 responses for which the recorded duration exceeded 9 hours indicating that a respondent had neglected to exit Qualtrics before submission.

a much more reasonable mean of 25.33 minutes and median of 19.43 minutes. User feedback was solicited via textbox; however, the responses were not analyzed at this stage of the research and will be discussed in Chapter 5. The complete list of usability comments can be seen in Appendix T.

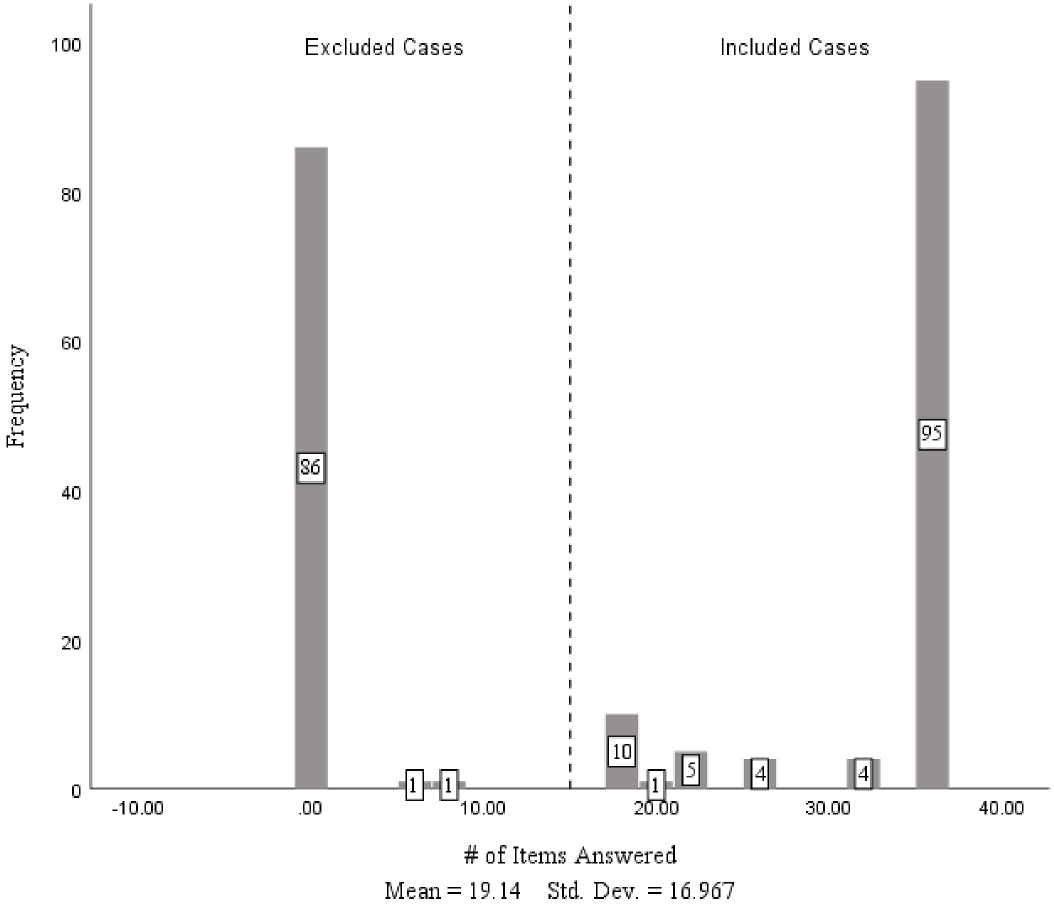


Figure 4.16. Histogram of Response Patterns

Based on the participant response, the data set was trimmed to eliminate responses that would be detrimental to the instrument validation purpose. All *Complete Responses* and *Complete Item Responses* were retained, as well as any *Partial Responses* for which at least 15 items had been answered. The justification for this cutpoint was based on a histogram (Figure 4.16) revealing a cluster of 20 respondents between the *Essentially Blank* (n = 88) and *Essentially Complete* (n = 99) response patterns. Unlike in Phase IIIA where the paucity of response patterns justified removal of incomplete responses, the relative contribution to total sample size of this group was too substantial to ignore and thus these respondents were retained for the partial information that would be contributed. Furthermore, with no reliable means of determination if these uncompleted items represented data that was missing at random, there was no compelling justification for removal. Ultimately upon elimination of appropriate partial responses (n = 88) and Master's respondents (n = 7), the trimmed dataset consisted of 112 responses. This is the data used in all subsequent analyses.

4.1.3.8 Phase IIIB Items Data

Analogous to the analysis of Phase IIIA, a cursory glance at respondent performance was conducted to verify that scores were not inappropriately high or low relative to expectation. In this instance, overall respondent performance on the instrument suggested moderate levels of *p*-value fluency with total scores indicating respondents correctly answered slightly better than half of the items, on average – scores that are commensurate with those of the local field test participants. Specific scoring details, in total and by sub-section, can be seen in Table 4.20 and in Figure 4.17.

Table 4.20 – Phase IIIB Scoring Statistics

Descriptive Statistics					
Instrument Section	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
P/CP Score	112	0.222	0.944	.60962	0.152183
RV Score	112	0.000	0.944	.50942	.262798
Total Score	112	0.111	0.889	.55952	.165321

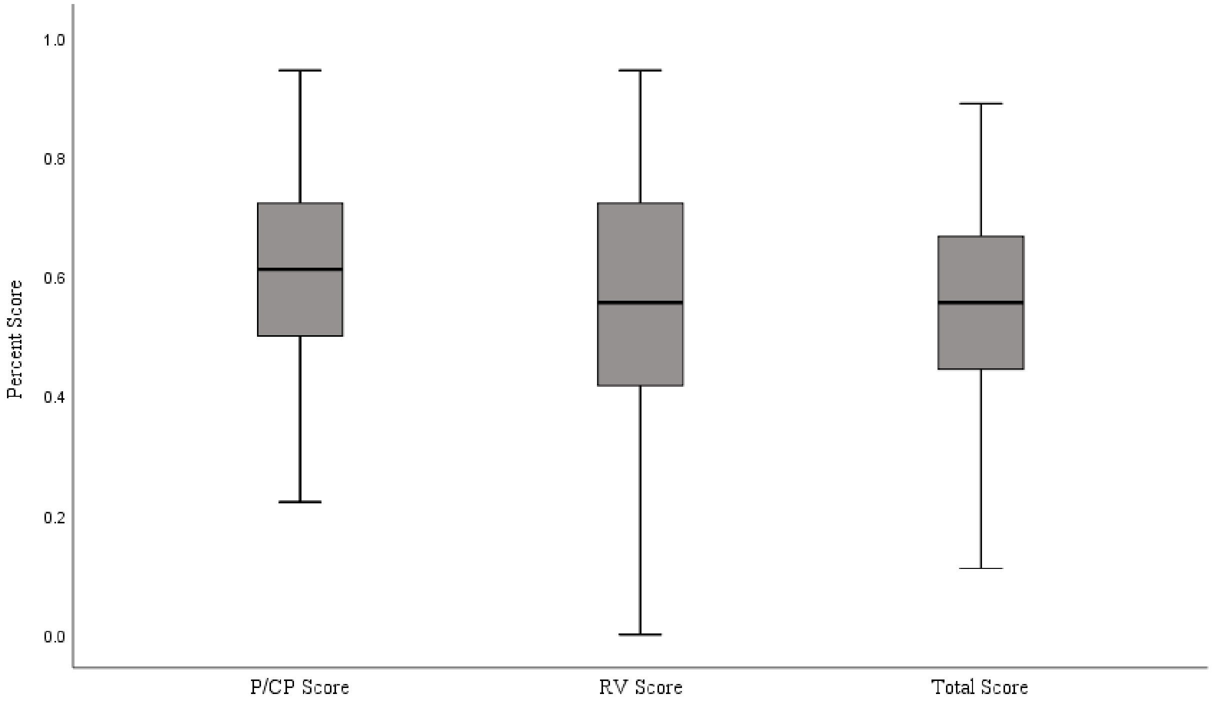


Figure 4.17. Phase IIIB Distribution of Scores by Section

Turning our attention to the *item performance*, this was addressed on the whole and via subsection. Considering the P/CP section first, Figure 4.18 shows the percentage of correct answers by item arranged in numerical order. Items that received more than 84 correct responses (n = 5) or less than 28 correct responses were flagged for consideration, but were not eliminated solely on this criterion at this stage of the analysis.

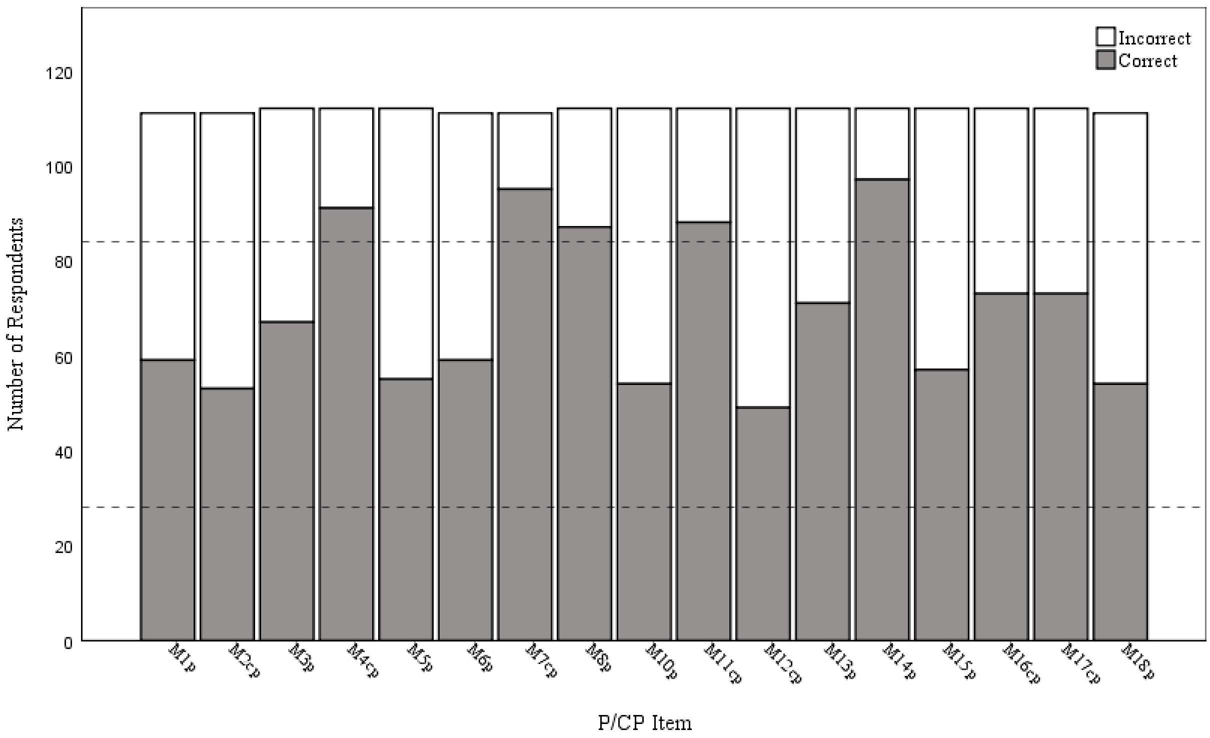


Figure 4.18. Participant Performance – P/CP Section

A similar examination of the research vignette items is displayed in Figure 4.19, arranged in numerical order. There were no items that were flagged for being too *hard* (i.e., < than 28 correct responses), nor were any flagged for potentially being too *easy* (i.e., > 84 correct responses).[25]

[25] The demarcations of y = 28 and y = 84 were decided based on a 25%/75% success rate. The researcher acknowledges that with the variation in number of responses, these thresholds are not equally stringent across items; however, on the basis that missing responses are no different than incorrect responses (the most conservative position is the assumption that ‘blank = unknown’), the success percentages would fluctuate but the characterization of ‘easy’ and ‘hard’ would not for the current data and thus were not adjusted.

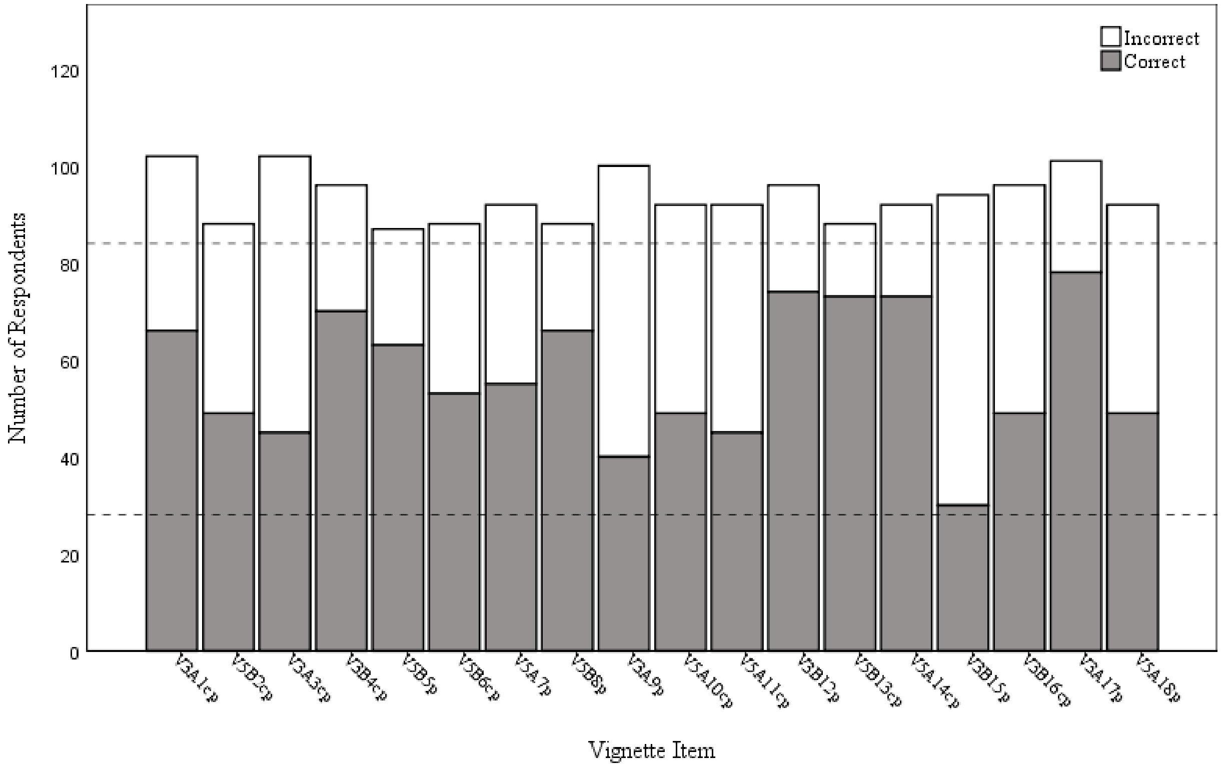


Figure 4.19. Participant Performance – Vignette Section

Analogous to the analysis of Phase IIIA, the internal consistency of the items across formats was investigated. Pearson correlations were computed to measure the strength of the relationship of each subsection score with the total score (see Table 4.21; cells display *r* with the *p*-value in parentheses). In this assessment, high positive correlations (i.e., close to 1.0) were desired as this would indicate that higher scores on a given subsection correlates with higher scores overall.

Table 4.21 – Correlation Matrix – Phase IIIB

Correlations

	Total Score	P/CP Score	RV Score	Experience
Total Score				
P/CP Score	.630 (< .001)			
RV Score	.893 (< .001)	.214 (.024)		
Experience	.337 (.002)	.277 (.012)	.317 (.004)	
Status	.007 (.949)	-.001 (.995)	.013 (.908)	-.139 (.217)

The research vignettes, as is consistent with previous phases of the study, showed the stronger correlation with total performance as compared with that of the P/CP section; however, it is worth nothing that both correlations are a bit weaker than those seen in the local field test.

Also in keeping with the Phase IIIA analysis, an investigation of the correlation between scores and coursework as a proxy for convergent validity was performed. In theory, the expectation is that training improves fluency and thus high, positive correlations would be desired for this measure. In reality, neither the total score (r = .337, p = .002), nor either subsection score (r = .277, p =.012; r = .317, p = .004), correlated satisfactorily with this measure; the indicated relationship here is weaker than that of Phase IIIA (after the subpopulation correction had been applied). That the correlations are positive (albeit low) is somewhat reassuring when taken at face value because it implies that the *more coursework = higher scores* relationship is maintained; however, the marginal contribution of additional coursework implied by this analysis again suggests that either fluency does not improve with additional training or that coursework is an inappropriate proxy for a convergent validity measure.

Linear regression was employed to model the relationship between courses completed and total score, merely as a means of comparison with the Phase IIIA results (not based on the expectation that this would be necessarily meaningful in and of itself). The suggested relationship ($\beta_0 = .523, SE_{\beta_0} = .029;\ \beta_1 = .035, SE_{\beta_1} = .011$; ANOVA F = 10.273; p = .002) indicated that a person with no experience would score about 52% on average with an expected increase of about 3% for each additional course completed – nearly 1/3 of the per course contribution suggested in the pilot test, but a figure commensurate to that seen in Phase IIIA before the Master’s students had been filtered. See Figure 4.20 for details.

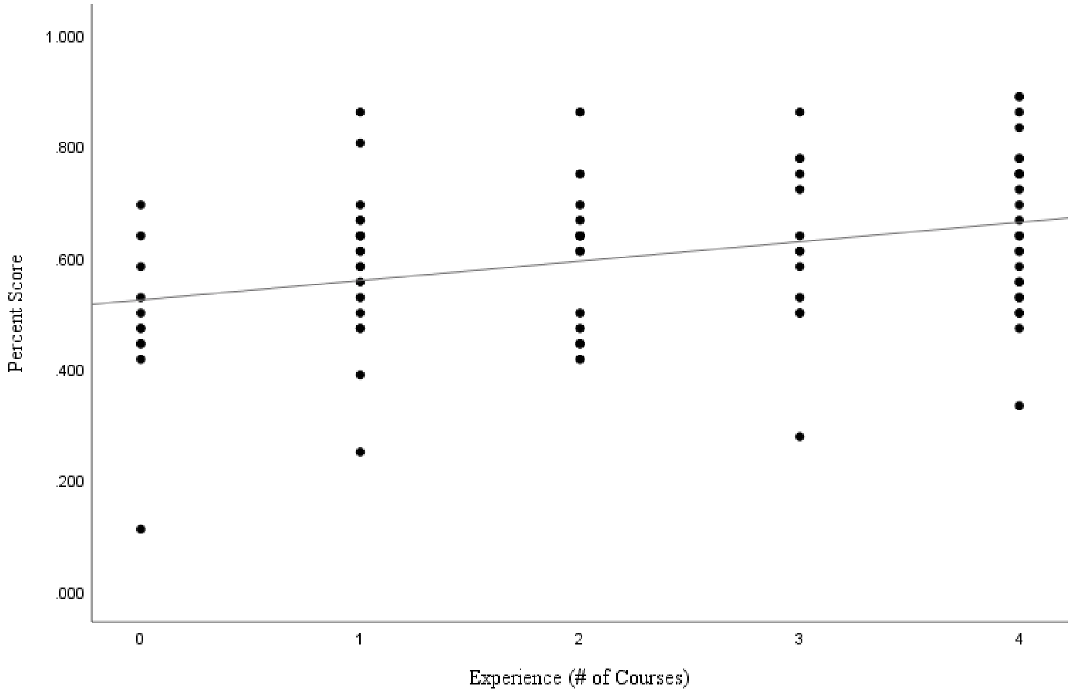


Figure 4.20. Regression Results – Phase IIIB

Again in accordance with analysis of data from previous phases of this research, the final assessment of instrument performance was a check on internal reliability, both in total and by subsection. When using Cronbach's alpha, the closer a value is to 1.0, the more reliable the measure. Conventional wisdom suggests that a value of 0.70 is the lowest acceptable threshold (Nunally, 1978). As can be seen in Table 4.22, this threshold has not been met by subsection for either the P/CP or Odd vignettes, but has been roughly attained for the instrument as a whole.

Analogous to the Phase IIIA analysis, follow-up reliability diagnostics were utilized to detect which, if any, items are problematic and adversely affecting the instrument's reliability. These measures included the *Corrected Item-Total Correlation* (do not wish to see negative values for this measure) and *Cronbach's Alpha if Item Deleted* (do not wish to see values that exceed the global alpha) measures. In this analysis, items were classified as *poorly functioning* if the *Corrected Item-Total Correlations* fell below the 0.05 threshold. This criterion identified six potential issues with the P/CP items and two with the vignettes (details can be found in Appendix U).

Table 4.22 – Cronbach's Alpha Data Phase IIIB

Subset	# of Items	Alpha	Diagnosis	Items
P/CP	18	0.509	6 items are poorly functioning	M12, M18, M14, M11, M2, M10
RV odd	18	0.609	2 items are poorly functioning	V3B15, V5B5
All items	36	0.718		

4.1.3.9 Field Test Data Consolidation

The original intention of this research was to use the local field test (Phase IIIA) for instrument validation and to use the national sample (Phase IIIB) in pursuit of the second research question. Upon consideration of the size and nature of the Phase IIIA dataset (i.e., initial *n* of 79 subsequently reduced to 35), the decision was made to postpone instrument validation until Phase IIIB data was in hand. Had the national field test produced responses on the order of 400 as planned, that dataset would have been sufficiently large for the investigation of test score estimation and other related validation measures and, accordingly, the use of the local field test data could have been discontinued at that time. Any insights gleaned in the way of usability would inform the next iteration of the instrument, but there would have been no need to utilize that data for any further analyses than had already been performed. Unfortunately, the response rate was much lower than anticipated and the national field test yielded a working data set of only *n* = 112 responses – a magnitude just large enough to justify exclusion of the Phase IIIA data, but just small enough not to dismiss that notion outright. As stated previously in this chapter, the research team decided to postpone certain aspects of the validation plan until both sets of data (i.e., Phases IIIA & IIIB) were in hand and to pool them when appropriate and/or necessary. Having reasonably met the *necessary* threshold, the following section will describe how the research team assessed the *appropriateness* criterion.

The goal of the national sample recruitment was to assess the *p*-value fluency of doctoral students enrolled at R1 institutions. Virginia Tech, being an R1 institution, certainly meets this criterion (thus justifying inclusion of the local field test data); however, due to the nature of the recruitment, there was reason to believe that this sample might be qualitatively different than the national sample somehow and thus might exert undue influence on the results due to its cluster size (i.e., *n* = 33 vs. cluster sizes of 4 or 8 or 10, etc., of the other institutions). The decision to pool the data (or not) was based on a three-part investigation: Are there demographic differences? Are there performance differences? Are there measurement differences?

The investigation into demographic similarities began with a look at the Race/Ethnicity and Sex/Gender variables. Tables 4.23 and 4.24 reveal that both samples were predominantly White but that the distribution amongst the other race categories was somewhat dissimilar. The Sex/Gender distributions were very nearly reversed: the national sample was heavily female (69.1%) while the VT sample was primarily male (68.6%).

Table 4.23 – Demographic Comparison of Field Test Data (By Percent)

		Data Source	
Demographic Classification		**National Sample**	**VT Sample**
Sex/Gender	Male	30.90%	68.60%
	Female	69.10%	31.40%
Race/Ethnicity	White/Caucasian	67.90%	39.40%
	Asian	13.60%	36.40%
	Black/African	1.20%	0%
	Hispanic/Latino	4.90%	12.10%
	Other/Mixed	12.30%	12.10%

Table 4.24 – Demographic Comparison of Field Test Data (By Count)

		Data Source	
Race/Ethnicity	**Sex/Gender**	**National Sample**	**VT Sample**
White/Caucasian	Male	20	8
	Female	35	5
Asian	Male	2	10
	Female	9	2
Black/African	Male	0	0
	Female	1	0
Hispanic/Latino	Male	1	3
	Female	3	1
Mixed/Other	Male	2	3
	Female	8	1
Declined to Provide	Male	0	0
	Female	0	2
	Declined to Provide	31	0
	Totals	112	35

Looking beyond the ubiquitous demographic markers (of race and sex), the datasets were compared on the basis of student status and experience (Table 4.25). Here we notice that the national sample has a fairly uniform distribution across Status (i.e., year in school) with nearly as many 5th year students as first year. This pattern was in contrast to the heavily right-skewed data of the VT sample that was heavy on the first- and second-years. Multicollinearity between status and course-taking behavior (i.e., more senior students are likely to have completed more courses) is likely, but worth investigating regardless. Here we notice that there is no discernable pattern in either dataset: there is a *slight* majority in the national sample of students who have completed more than three courses and a *slight* majority of students in the VT sample who have taken two courses, but all categories seem to be healthily represented.

Table 4.25 – Demographic Comparison of Field Test Data by Experience

		Data Source	
Experience Marker		**National Sample**	**VT Sample**
Status	1st yr PhD	23.50%	45.70%
	2nd yr PhD	16.00%	34.30%
	3rd yr PhD	16.00%	11.40%
	4th yr PhD	16.00%	5.70%
	5th yr PhD	21.00%	3%
	Other (Not Master's)	7.40%	0.00%
# of Required Courses	None	22.00%	38.20%
	One	26.80%	11.80%
	Two	20.70%	35.30%
	Three	20.70%	2.90%
	More than Three	9.80%	11.80%
# of Courses Taken	None	13.40%	0%
	One	23.20%	28.60%
	Two	17.10%	31.40%
	Three	15.90%	22.90%
	More than Three	30.50%	17.10%

An interesting follow-up to the previous investigation was to look not at *total* courses completed, but rather the *discrepancy* between those required and those completed. In other words, are the students agreeing to participate in this research somehow more inclined to *p*-value fluency as a result of interest or experience and is this consistent across samples? Figure 4.21 shows the overlaid distributions of Course Discrepancy. We notice that while both samples tend to overachieve on average (take more coursework in statistical methods than is required), this characteristic is stronger in the VT sample than nationally. This is a somewhat unexpected result given the nature of the recruitment in that intuition might assume that the national sample students, offered nothing more than a potential raffle prize, would be more likely to have a particular affinity for this topic (and thus more likely to participate) than the VT students for whom extra credit was granted.

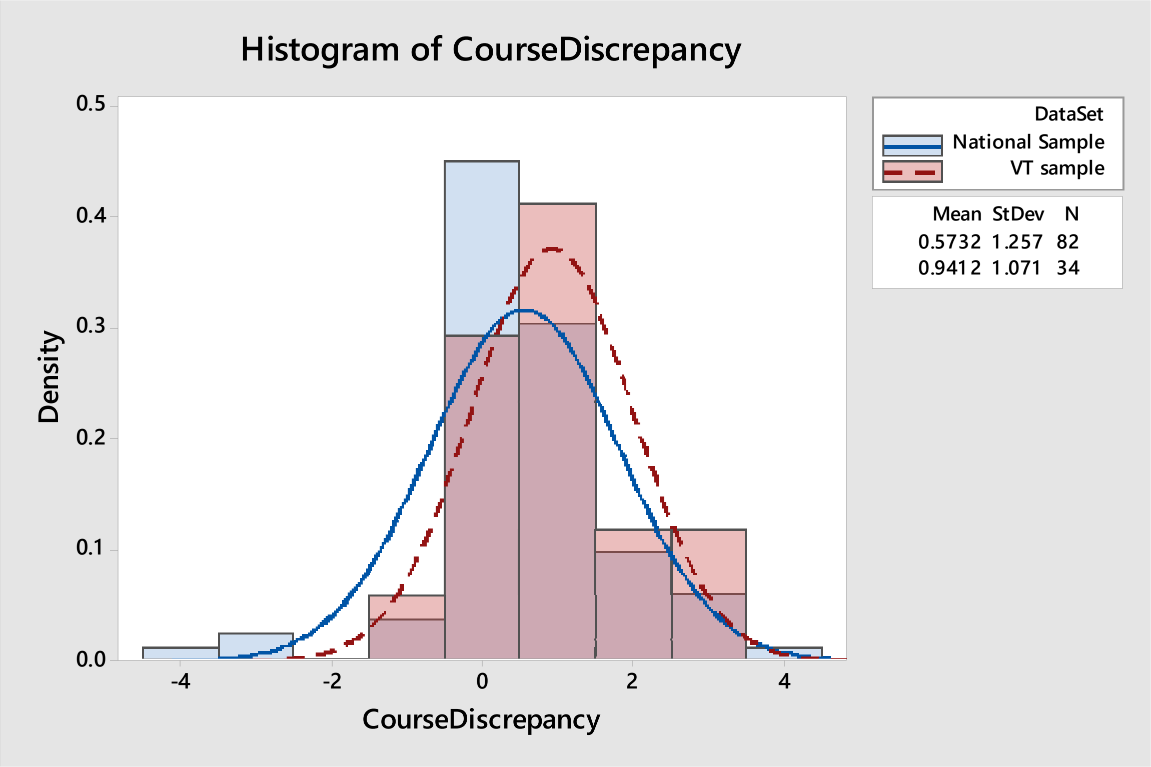


Figure 4.21. Histogram of Course Discrepancy by Sample

In summary, the most striking differences between the local and national samples were by Sex and Status. The VT sample was more inexperienced (more 1st-year students) and more male; however, neither of these are necessarily grounds for dismissal of the notion of a pooled dataset. Natural fluctuations in sex and race differences exist from institution to institution and there is no reason to expect that one university examined singly would match the aggregate distribution of a collection of institutions. Any of the participating institutions considered individually would likely also show deviations from the aggregate in the same way that VT has done. The status differences between the samples, on the other hand, could possibly be problematic if 1st-year students (for instance) engage with the instrument differently than more advanced students in ways that are antithetical to the validation process (i.e., students with less experience *should* theoretically score lower on this assessment and thus the infusion of these students in the mix would not necessarily undermine the validation efforts unless their performance contradicted reason or expectation.); however, this is inestimable at the present time and will be revisited after the performance and measurement subsections of this analysis.

The investigation into performance similarities began with a look at the distribution of scores across the subsections (P/CP, Vignettes) and in total (see Figures 4.22, 4.23, & 4.24). The national sample performed better on the P/CP subsection (61% vs. 53%) on average with nearly identical standard deviations. This trend was reversed on the Vignette subsection with the VT sample slightly outscoring the national sample (55% vs. 51%) on average. It is worth noting, however, that the national sample had an unusually large group of 0[26] scores that are clearly influencing the mean calculations and absent that, there would likely be no performance difference between the groups. Overall, the national sample and VT samples were nearly identical in mean and standard deviation, with a slight edge going to the national sample (54% vs. 56%).

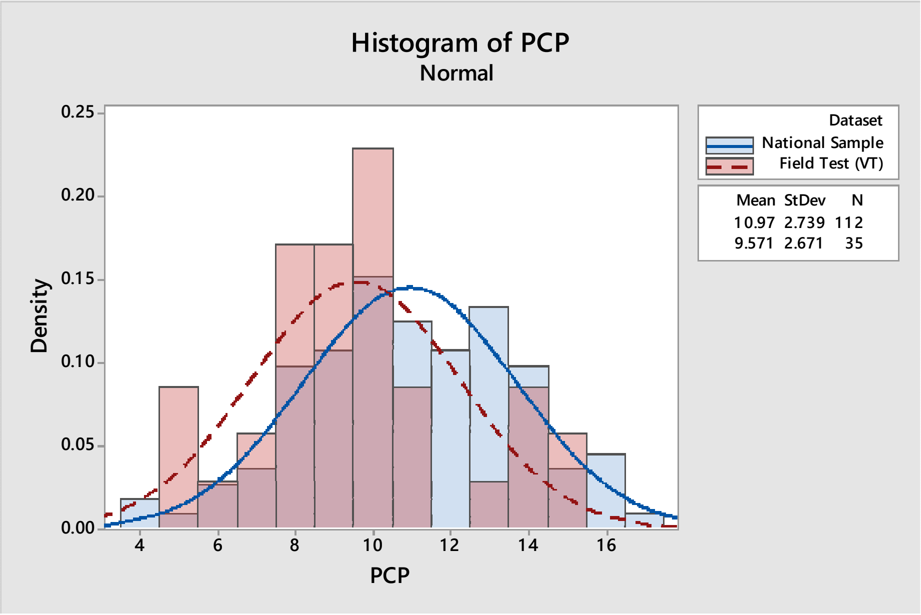


Figure 4.22. Distribution of P/CP Scores by Sample

[26] Note that the national sample allowed for the inclusion of incomplete/partial responses and the VT sample did not. The ordering of the sections (P/CP first, RV last) would contribute to the existence of 0 scores on the RV section due to survey attrition.

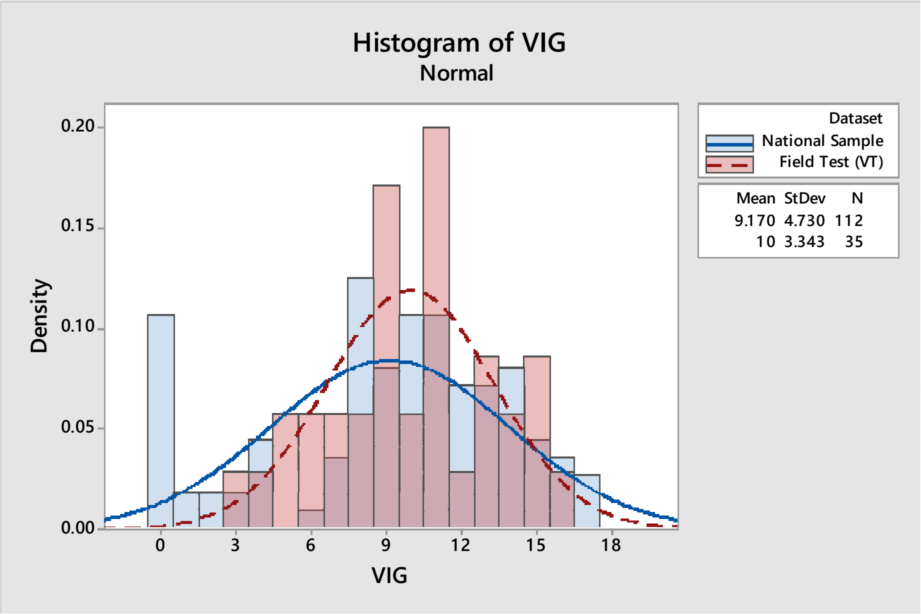


Figure 4.23. Distribution of Vignette Scores by Sample

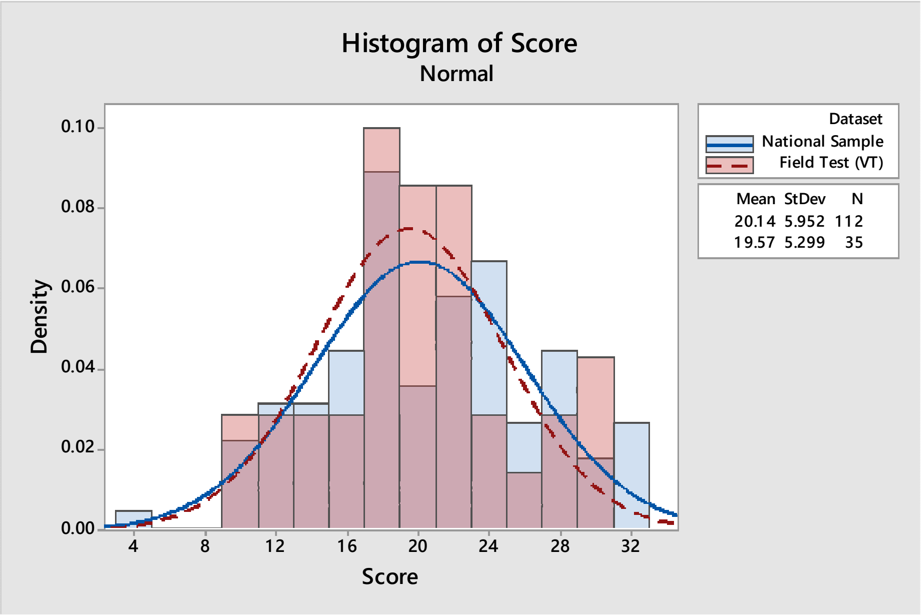


Figure 4.24. Distribution of Total Score by Sample

Independent samples t-tests (Table 4.26) suggest that the difference on the P/CP subsection is statistically significant (p = .009), but that the discrepancies on the Vignette subsection and in overall performance are not (p = .253, p = .591 respectively). The moderate effect size of the P/CP difference (Cohen's *d* = .52) is worth noting; however, this must be kept in perspective of the units of measurement of the instrument. Is it probably unreasonable to assume that there is an appreciable difference in the *p*-value fluency of individuals differing by one (or two) item(s) on this assessment.

Table 4.26 – Independent Samples t-Test Comparison of Performance by Sample

Section	Data	n	Mean	SD	t	df	p	Mean Diff	Std Error	95% CI
P/CP	NS	112	10.97	2.739	2.694	58.1	0.009	1.402	0.527	(0.36, 2.44)
	VT	35	9.57	2.671						
RV	NS	112	9.17	4.73	-1.152	80.234	0.253	-0.83	0.72	(-2.26, .603)
	VT	35	10	3.343						
Overall	NS	112	20.14	5.952	0.54	63.091	0.591	0.571	1.058	(-1.54, 2.69)
	VT	35	19.57	5.299						

In summary, there is no compelling evidence to suggest that the performance of the VT sample is alarmingly distinct from that of the national sample. Furthermore, small mean differences are not necessarily indicative of lack of invariance. The VT students could reasonably be above or below the national average depending on the institutions participating, but is in any case a viable subgroup of the population of doctoral students nationally.

The investigation[27] into measurement similarities began with a classical item analysis. Using jMetrik software, the item difficulty, standard deviation, and discrimination were computed for each of the 36 items. This analysis was conducted separately for the VT sample data and the national sample data as a means of comparing estimates therein.

Item difficulty is the mean item score and in the case of binary items, it represents the proportion of examinees who answered correctly. The expectation is that item difficulties should be as close to 0.5 as possible (for binary items) as this value "maximizes item variance and subsequently increases score reliability" (Meyer, 2014, p.44). Acknowledging that every item will not achieve the ideal value, Allen and Yen (1979, p.121) recommend a range of acceptable

[27] Selected results for the tests referenced in this section are presented in the text. Details for all tests can be found in Appendix V.

values between 0.3 and 0.7. Given the preliminary nature of this research and the objective of the current analysis (i.e., pending decision as to the appropriateness of a pooled data set), this threshold was expanded slightly to classify items as being *too easy* if difficulty values exceeded 0.75 and *too hard* if difficulty values fell below 0.25. This criterion identified three items as being unacceptably low/high in the VT sample and five in the national sample.

Item discrimination is the "extent to which an item differentiates between examinees that obtain different scores on the test" (Meyer, 2014, p.41). Higher values are desirable here because it indicates the item's ability to differentiate between respondents with similar but not identical scores; lower values indicate that the item can only differentiate between examinees of very different ability. Item discrimination values, being that they represent correlations between the item score and the total score, range from -1 to 1; however, only positive values are considered acceptable. Negative discrimination values would indicate that lower-scoring examinees answer this question correctly and would suggest that the item is flawed. Two types of item discrimination indices can be computed – Pearson or poly-serial – but standard practice is to use only one based on the nature of the data. (In the case of binary items, these are referred to as biserial and point-biserial correlations.) In this instance, the biserial option was selected for its reputed stability across different groups of examinees (Meyer, p. 42) since the participants in the field test represented a wide range of Status and Experience. Items having negative discriminations were flagged and any item having a positive discrimination in the range of 0.0 – 0.20[28] was deemed as having unsatisfactory discrimination. This criterion identified 6/3 negative items and 5/7 items as having potentially[29] unsatisfactory discrimination for the VT/national samples, respectively. Table 4.27 presents the difficulty and discrimination measures by item for

[28] This cutoff was adopted on the guidance given by Prometric (2017).

[29] Item discriminations are notoriously unstable, so these are considered *potentially* unsatisfactory to reflect the tentativeness of the findings.

each of the field test samples as well as a presentation of the disposition of the comparison therein. The primary objective of this particular analysis was not to diagnose problem items *per se*, but rather to determine the suitability of pooling the data. To that end, consistency of classification across samples was investigated (also presented in Table 31). Confirmed conflicts were identified for 5 of the 36 items and potential conflicts were identified in an additional 4 items[30].

[30] *Confirmed* conflicts were classified on the basis that one sample identified the item as problematic and the other did not. *Possible* conflicts were classified as such based on one sample identifying a minor issue that the other sample did not confirm. *Agreement* was declared when the item was either satisfactory or problematic in both samples.

Table 4.27 – Item Analysis by Sample

	VT Sample Results			National Sample Results			Comparison
Item	**Difficulty**	**Discrimination**	**Comments**	**Difficulty**	**Discrimination**	**Comments**	**Disposition**
M12	0.8	-0.2246	Negative discrimination	0.4375	-0.1878	Negative discrimination	agreement
M3	0.3429	0.7444		0.5982	0.2646		agreement
M12	0.3429	0.3565		0.5268	0.2184		agreement
M13	0.3714	0.2276		0.6339	0.0721	Discrimination too low	possible conflict
M18	0.6	0.0801	Discrimination too low	0.4821	0.0814	Discrimination too low	agreement
M9	0.5429	0.0341	Discrimination too low	0.4196	0.2719		possible conflict
M4	0.5143	0.4713		0.8125	0.3764	Item too easy	possible conflict
M16	0.6	0.6352		0.6518	0.2541		agreement
M14	0.8286	-0.0975	Item too easy / Negative discrimination	0.8661	0.1487	Item too easy / Discrimination too low	agreement
M6	0.4571	0.0441	Discrimination too low	0.5268	0.16	Discrimination too low	agreement
M11	0.7429	0.2149		0.7857	-0.1513	Item too easy / Negative discrimination	conflict
M7	0.6	0.3594		0.8482	0.4314	Item too easy	possible conflict
M5	0.2286	0.1143	Item too hard / Discrimination too low	0.4911	0.0692	Discrimination too low	agreement
M2	0.4286	-0.1951	Negative discrimination	0.4732	0.0976	Discrimination too low	agreement
M10	0.3143	0.1276	Discrimination too low	0.4821	0.1391	Discrimination too low	agreement
M15	0.5714	0.4191		0.5089	0.337		agreement
M8	0.7429	0.3205		0.7768	0.3144	Item too easy	agreement
M17	0.5429	0.4067		0.6518	0.2795		agreement

VT Sample Results				National Sample Results			Comparison
Item	**Difficulty**	**Discrimination**	**Comments**	**Difficulty**	**Discrimination**	**Comments**	**Disposition**
V3A1	0.6571	0.4469		0.5893	0.5669		agreement
V3A3	0.3714	0.2879		0.4018	0.367		agreement
V3A9	0.4	-0.2293	Negative discrimination	0.3571	0.4958		conflict
V3A17	0.6571	0.4788		0.6964	0.6245		agreement
V3B4	0.5429	0.348		0.625	0.6604		agreement
V3B12	0.9429	-0.3669	Item too easy / Negative discrimination	0.6607	0.469		conflict
V3B15	0.5429	0.5258		0.2679	-0.1242	Negative discrimination	conflict
V3B16	0.3714	0.6779		0.4375	0.4936		agreement
V5A7	0.4	0.3895		0.4911	0.7128		agreement
V5A10	0.4857	0.362		0.4375	0.5347		agreement
V5A11	0.4286	0.5071		0.4018	0.3999		agreement
V5A14	0.7714	-0.1724	Item too easy / Negative discrimination	0.6518	0.6391		conflict
V5A18	0.5143	0.3683		0.4375	0.4404		agreement
V5B2	0.5714	0.464		0.4375	0.5553		agreement
V5B5	0.4286	0.5374		0.5625	0.4242		agreement
V5B6	0.5429	0.2609		0.4732	0.649		agreement
V5B8	0.6571	0.3363		0.5893	0.7366		agreement
V5B13	0.7143	0.528		0.6518	0.8366		agreement

The investigation of measurement similarities continued with a reliability analysis. Reliability, according to Meyer (2014), “refers to the reproducibility of test scores” (p.53). The jMetrik software calculates a variety of reliability measures, but we focus on only one in this research: *Coefficient Alpha*. This measure is derived from the early work of Kuder and Richardson who developed two indices designed to assess internal consistency of an instrument, so-named the *KR20* and *KR21*[31]. The primary distinction between the two being that the latter makes the assumption that all items are equally difficult (an untenable assumption in this instance). A lower-bound estimate on the reliability of an instrument, α is a function of the number of items in a test, the average covariance between item-pairs, and the variance of the total score (Cronbach, 1951). Both samples scored reasonably well on this measure (*Coefficient alpha* = .7732 w/ *SEM* = 2.7575 vs. .7269 w/*SEM* = 2.7692) and the measure was notably consistent across samples.

The next investigation conducted was for DIF – differential item functioning. DIF “occurs when one group of examinees has a different expected item score than comparable examinees from another group [indicating] that an item is measuring something beyond the intended construct and is contributing to construct irrelevant variance” (Meyer, p. 69). Distinct from *item impact* which investigates whether performance differences are a result of group differences on the measured trait, DIF controls for group differences by “evaluating the performance of comparable focal and reference group members” (p. 69). Using the national sample as the reference group and the VT sample as the focal group, DIF analysis was performed in jMetrik as an additional check on the feasibility of pooling the data. Indication of DIF on

[31] Both the *KR20*and *KR21* are designed for use with binary items. Subsequent work by Guttman led to the *Coefficient alpha* measure generalizing the *KR20* to polytomous items. In the case of binary items, the result is the same. jMetrik refers to this as *Coefficient alpha* in the output table and so this paper will use that nomenclature for consistency.

more than a couple items would suggest that the instrument is not functioning in equivalent ways for the two samples and this would call into question the invariance.

DIF analysis can be conducted in a variety of ways. jMetrik provides the Mantel-Haenszel chi-square procedure, the common odds ratio effect size, and the ETS DIF classification levels. The common odds ratio effect size ranges from [0, ∞) with an expected value of 1; the ETS Delta option transforms the common odds ratio into a symmetric measure, centered about 0, with range [-4, 4][32]. For this research, the *ETS* option was employed.

Using the *ETS* option in jMetrik, three levels of severity of DIF (A = negligible, B = moderate, C = large) are assigned based on the magnitude of the [transformed] common odds ratio (Meyer, p. 75)[33]. When the focus is instrument and item development, classifications of B or C would be cause for concern and would justify revision or removal of the item. The present analysis was designed to measure instrument performance invariance across samples and thus no item modifications were performed at this time, but instead, severe DIF classifications were taken as evidence against the suitability of a pooled data set. Two items were assigned C-level DIF ratings and six were assigned B ratings – roughly 22% of the total items (See Appendix V for details). Keeping in mind that DIF estimates are sensitive to inequality of focal and reference sample sizes, and considering that some of the DIF items were already identified as problematic on other measures (e.g. item discrimination), this is not necessarily an indication that the data should not be combined.

The final investigation into measurement similarities investigated dimensionality and IRT (item response theory) model selection. Selection of the appropriate IRT model is contingent

[32] More information on jMetrik and specifics about the algorithms can be found at: https://itemanalysis.com/

[33] A item: Chi-square *p*-value > 0.05 or 0.65 < *COR* < 1.53; B item: not an A or C item; C item: *COR* < 0.53 and the upper bound of the 95% CI is < 0.65 or *COR* > 1.89 and the lower bound of the 95% CI is > 1.53. (https://itemanalysis.com/faq/what-is-the-information-in-the-dif-analysis-output/)

upon determination of the dimensionality of the latent space, and thus that must be ascertained prior to model specification. This research had assumed the position of unidimensionality thus far, but that assumption will be tested in this investigation.

Two options exist for determining dimensionality: EFA (exploratory factor analysis) and CFA (confirmatory factor analysis). In the former method, the researcher specifies *a priori* the number of latent dimensions and the algorithm generates factor loadings for each item representing the relative contribution each item is having on that particular factor. It is up to the researcher to assign meaningful labels to the factors after the item assignment has been declared. In the latter method, the researcher applies a theoretical framework to dictate the assignment of items to factors and the algorithm generates factor loadings therein. The post hoc decision making in EFA consists of matching the loadings to the content of the items and determining if the underlying structure suggested by the model makes sense. The post hoc decision making in CFA consists of assessing the model diagnostics and parameter estimates as being evidence for/against the proposed factor structure.

In assessing the measurement invariance between the VT and national field test samples, factor analysis was conducted separately on each sample. Fixing the IRT model to be 2PL[34] across all iterations, exploratory factor analysis was conducted with one, two, and three factors specified and a confirmatory factor analysis was conducted with 2 factors specified (P/CP on Factor 1, Vignettes on Factor 2), as well as a Bifactor model with the same two sub-factors specified as in the CFA. For each of the five proposed structures, the AIC, BIC, and *-2log likelihood* measures were recorded (Table 4.28).

[34] The 2PL model is considered the default for dichotomous items. The suitability of this model to this data will be assessed in a later analysis.

AIC – Akaike Information Criterion – is an estimate of the relative quality of a statistical model that estimates the relative amount of information lost by a given model while attempting to balance both over- and under-fitting concerns (i.e., the tradeoff between goodness of fit and model simplicity). BIC – Bayesian Information Criterion – is another estimate of the relative quality of a statistical model and is closely related to the AIC. In both measures, the calculation is dependent upon the number of parameters being estimated, but in the case of the BIC, the penalty term is larger. In both formulas, *-2 log likelihood* features prominently[35]: $AIC = -2\log(L) + 2p$; $BIC = -2\log(L) + p\log(N)$, but impose a penalty term (in *all* cases, the smaller the measure, the better the model). It is permissible to simply compare *-2log likelihood* values by themselves, but generally AIC/BIC measures are considered the preferred estimates as they are less susceptible to sample size concerns. When comparing a model set using all three criterion, the expectation (or at least, hope) is that all three measures will indicate the same *best* model; however, when the measures differ, the AIC will prefer the more complicated model.

Table 4.28 – Factor Analysis Results by Sample

VT Sample				**National Sample**			
Model	**-2LogLikelihood**	**AIC**	**BIC**	**Model**	**-2LogLikelihood**	**AIC**	**BIC**
Bifactor	1426.16	1642.16	1810.14	Bifactor	4239.94	4455.94	4749.54
EFA1(2PL)	1489.7	1633.7	1745.69	EFA1(2PL)	4337.21	4481.21	4676.94
CFA2(2PL)	1508.39	1652.39	1764.37	CFA2(2PL)	4390.01	4534.01	4729.74
EFA3(2PL)	2287.34	2569.34	2788.64	EFA3(2PL)	4202.71	4484.71	4868.02
EFA2(2PL)	2341.68	2555.68	2722.1	EFA2(2PL)	4256.94	4470.94	4761.82

Calculations were performed using the *Multidimensional IRT* command in IRTPRO software package. The results[36] indicated that the most likely factor structure for the VT data was either the Bifactor model or the EFA1 model. For the national sample, the results were a bit

[35] Formulas from: http://support.sas.com/documentation/cdl/en/etsug/63348/HTML/default/viewer.htm#etsug_severity_sect022.htm

[36] Full IRTPRO output in Appendix W for all candidate models.

inconclusive as each of the three criteria identified a different *best* structure: Bifactor, EFA1, and EFA3.

In both samples, there was evidence to suggest that a unidimensional[37] structure might be feasible and thus that became the working assumption in the next analysis. To examine invariance of IRT model across samples, the following models were considered: Rasch, 1PL, 2PL, and 3PL. The 2PL is the default model choice for dichotomous items. In this model, the following two parameters are estimated: item difficulty (β) and item discrimination (α). In the 3PL model, an additional parameter for guessing (*g*) is estimated. Both the Rasch and the 1PL model are special cases of the 2PL model where the item discrimination is fixed across items. In the former, $\alpha = 1.0$, and in the latter, α is estimated as a function of the available data. Calculations were performed using the *Unidimensional IRT* command in IRTPRO software[38]. Parameter estimates were calculated for each of the four models and selection was based on the following criteria (see Table 4.29): standard errors, goodness-of-fit statistics, and LID (local item dependence). For the national sample, the diagnostics suggest that the 2PL model is a reasonable choice. For the VT sample, the results are less conclusive. The 2PL model is certainly viable, but so also is the 1PL and possibly the Rasch. As the sample sizes are small, these results are considered tentative and warrant further investigation.

In summary, the measurement comparison is for all intents and purposes inconclusive. While small differences were detected for item discrimination, this did not hold when considering item difficulty and reliability measures. In terms of dimensionality and model

[37] Both the EFA1 and Bifactor models were candidates for selection. The EFA1 model loads all items onto a single factor and in the Bifactor model, the assumption is that all items load onto a single main factor and then separate out subsidiarily into subfactors. Either way, the notion of one underlying factor is supported.

[38] Full IRTPRO output in Appendix X for all candidate models.

selection, there was evidence suggesting that a unidimensional 2PL model would be appropriate for both data sets; however, the data was not compelling enough to disregard other options.

Table 4.29 – IRT Model Diagnostics by Sample

National Sample						
Model	**S.E.**	**χ2 Item Level Diag.**	**LID**	**-2LogLikelihood**	**AIC**	**BIC**
Rasch	0 items	5 items	7 items	4521.93	4593.93	4691.79
1PL	0 items	6 items	6 items	4491.3	4565.3	4665.88
2PL	3 items	3 items	3 items	4336.84	4480.84	4676.57
3PL	2 items	6 items	2 items	4362.18	4578.18	4871.78
VT Sample						
Model	**S.E.**	**χ2 Item Level Diag.**	**LID**	**-2LogLikelihood**	**AIC**	**BIC**
Rasch	0 items	4 items	0 items	1554.66	1626.66	1682.65
1PL	0 items	6 items	1 item	1546.21	1620.21	1677.76
2PL	2 items	2 items	0 items	1489.64	1633.64	1745.63
3PL	1 item	9 items	0 items	1511.04	1727.04	1895.02

The task of determining the invariance of the field test samples was pursued on the grounds that the Phase IIIA data was insufficient on its own for instrument validation (as was the original intention of this phase of the research) and that while the Phase IIIB data was likely sufficiently large enough to stand alone, the diagnostics and model parameter estimates might be enhanced when conducted on the larger, pooled dataset. Upon review of the diagnostics, neither position (pooled vs. not) is emphatically favored: the demographics were somewhat dissimilar, the performance statistics were rather similar, and the measurement indices (the tiebreaker) were essentially inconclusive. A summary of this investigation can be seen in Table 4.30. While a defensible argument could be made on either side, the evidence was not sufficiently compelling to dictate that the data should *not* be pooled and therefore the two field test samples were combined for the purposes of instrument validation. For this point forward, "the data" will refer to the pooled dataset consisting of n = 147 responses.

Table 4.30 – Summary of Invariance Investigation

Category	Indicator	Invariant?	Edge
Demographics	Race	Inconclusive	
	Gender	Yes	
	status	Yes	NS
	Coursework	Inconclusive	VT
Performance	P/CP section	Yes	NS
	RV section	No	
	Overall	No	
Measurement	Item Discrimination	Yes	NS
	Item Difficulty	No	
	Reliability	No	NS
	DIF	Inconclusive	
	Factor Structure	Inconclusive	
	IRT Model	Inconclusive	

4.1.4 Instrument Validation

A principal objective of this research was to address the following question: *Can a sufficiently reliable and valid measure of p-value misinterpretations (in a research context) be constructed*? To that end, a validation plan for this instrument was devised in the tradition of the Messick (1989, 1995) 6-component framework. In this section, evidence collected in each of the following six categories will be presented: Content, Substantive, Structural, Generalizability, External, and Consequential.

4.1.4.1 Content Validity

Constituent to content validity are three main components. Content relevance refers to the applicability of the measured construct to the target population. Representativeness refers to the ability of the instrument to capture the depth and breadth of the proposed construct. Technical quality refers to the ability of the items to dependably elicit accurate representations of respondent capability.

As to content relevance, the target population for this instrument is PhD candidates, i.e., future researchers, for whom p-value fluency is essential. The research community demands that its participants possess fluency in both facets of the research process: conducting research and evaluating the research of others. When conducting research, the researcher must understand the valid implications of the methodology employed and limitations therein and be able to select from a myriad of possible conclusions to report the most appropriate for the data and the situation; furthermore, when reviewing research, the reviewer must argue for or against the current methodology or interpretations of others. It is for these reasons that the item types were developed as they were, i.e., context-free and with context (i.e., in the evaluation of researcher actions), in order to simulate authentic conditions.

As to representativeness, the instrument was designed to address an extensive, research-based list of practitioner misinterpretations. The scope of the assessment was determined by a panel of experts (Greenland et al.) in *p*-value research and further, the individual items were repeatedly subjected to review by academicians responsible for teaching research methods to graduate students. These subject-matter experts are familiar with both the types of knowledge necessary for this type of work and also the frequently misunderstood topics therein.

As to technical quality, this was assessed with a variety of indicators including item difficulty and item discrimination. This was performed in jMetrik with the *Classical Item Analysis* command (output in Table 4.31). Item difficulty, as measured by the proportion of respondents who answered a given question correctly, should ideally be close to .5; particularly low or high values provide the least amount of discrimination. Item difficulties were deemed acceptable if falling within the $0.3 \leq x \leq 0.7$ range. This criterion identified five items in the P/CP section (all of which were *too easy*) and one item in the Vignette section (also *too easy*). Item discrimination for criterion-referenced assessments should be positive; negative values indicate that the question is likely disproportionately deceptive or misleading to those respondents who are scoring at the high level of the construct. Items having negative discriminations were flagged and any item having a positive discrimination in the range of 0.0 – 0.20 was deemed as performing unsatisfactorily. This criterion identified eight items in the P/CP section (two of which were negative, six which were positive but less than 0.2) and one in the Vignette section (positive, but near 0).

Table 4.31 – Classical Item Analysis Results

Item	Difficulty	Std. Dev.	Discrimination	Item	Difficulty	Std. Dev.	Discrimination
M12	0.5238	0.5011	-0.2056	V3A1	0.6054	0.4904	0.5361
M3	0.5374	0.5003	0.3539	V3A3	0.3496	0.4904	0.3508
M12	0.483	0.5014	0.2471	V3A9	0.3673	0.4837	0.3255
M13	0.5714	0.4966	0.1076	V3A17	0.6871	0.4653	0.5911
M18	0.5102	0.5016	0.0738	V3B4	0.6054	0.4904	0.5892
M9	0.449	0.4991	0.2099	V3B12	0.7279	0.4466	0.3495
M4	0.7415	0.4393	0.3775	V3B15	0.3333	0.473	0.0098
M16	0.6395	0.4818	0.3375	V3B16	0.4218	0.4955	0.5322
M14	0.8571	0.3511	0.09	V5A7	0.4694	0.5008	0.6436
M6	0.5102	0.5016	0.1371	V5A10	0.449	0.4991	0.4294
M11	0.7755	0.4187	-0.0644	V5A11	0.4082	0.4932	0.4203
M7	0.7891	0.4093	0.3868	V5A14	0.6803	0.468	0.4713
M5	0.4286	0.4966	0.0805	V5A18	0.4558	0.4997	0.418
M2	0.4626	0.5003	0.0358	V5B2	0.4694	0.5008	0.5209
M10	0.4422	0.4983	0.1401	V5B5	0.5306	0.5008	0.4475
M15	0.5238	0.5011	0.3495	V5B6	0.4898	0.5016	0.5575
M8	0.7687	0.4231	0.3161	V5B8	0.6054	0.4904	0.646
M17	0.6259	0.4856	0.3089	V5B13	0.6667	0.473	0.7672

4.1.4.2 Substantive Validity

A key component of substantive validity is that a theoretical framework must be established that supports the development of the construct map and operational definition from which the internal and external models must logically follow. This instrument has achieved this in the selection of Greenland et al.'s list of misinterpretations (i.e., the subset chosen for this research) as the theoretical framework. The items are a direct consequence (in fact, sometimes verbatim) of this basis. Furthermore, it has been established that this choice of theoretical framework supports both the internal and external models as specified.

P-value fluency, as defined by this research, is the ability to use *p*-values in a manner that is methodologically defensible and does not "contribute to the statistical distortion of the scientific literature" (as defined by Greenland, et al., 2016). A researcher with high *p*-value fluency is one who can appropriately interpret *p*-values, both in and out of context, and recognize when others have not. Inspired by the framework of Greenland and his colleagues (2016), a construct map was proposed (Figure 3.1) that identifies level of proficiency based on a respondent's ability to differentiate between misinterpretations and correct reporting of *p*-values.

Consistent with that construct map, this instrument was designed to assess the respondent at two process levels with the use of items that measured misinterpretations in a context-free setting and with items embedded in a research vignette setting. It was the speculative position of this research that the relative difficulty of the items was dependent upon the location within/without context, and that the former represents the more difficult task. Item hierarchy analysis of the field test data was used to investigate the strength of correlation between the item difficulty indices and the corresponding proposed process levels. These results suggest that the initial hypothesis was incorrect – the research vignette items (i.e., within context) were *easier* than the P/CP items (i.e., context-free) as can be seen in Figure 4.25. There was a small (but insignificant) negative correlation between the item type and difficulty (r = -.269, p = .112 with the context-free items as the reference group).

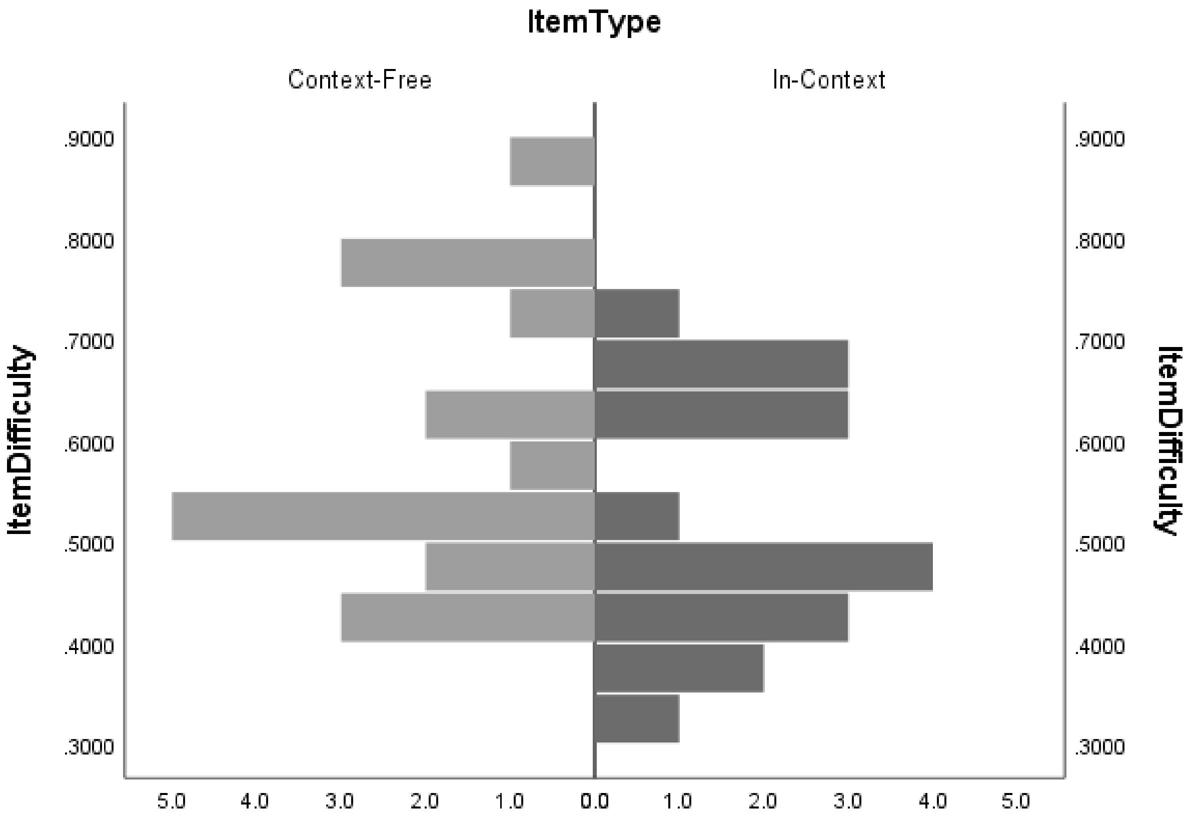


Figure 4.25. Distribution of Item Difficulties by Item Type

That the item hierarchy analysis did not corroborate the speculative process level assignment does not invalidate the instrument. The statistical insignificance of the negative correlation does not allow for a definitive conclusion; thus, while it is possible that the contextual items are easier for respondents and therefore require a lower level of *p*-value fluency to answer correctly, that determination cannot be made with the data in hand. Item revision would be the prudent course of action in the face of an established theoretical (or empirical) precedent for the difficulty of the contextual items, but absent such grounding, the more conservative approach is modification of the construct map and test blueprint instead – the presentation of which follows later in this section.

In addition to establishing a theoretical framework, substantive validity refers to the empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks. The data obtained from the cognitive labs and the usability study addressed substantive validity in this regard. Specifically, misleading and confusing words and phrases were removed from the item stems upon respondent identification therein. Furthermore, the decision to abandon the traditional T/F format in favor of the P/CP structure was in direct response to respondent feedback. Rather than cueing the participant to the nuances between a misinterpretation and a correct interpretation of a particular *p*-value concept, the items were instead provoking the respondent to focus on the "key word" falsifying the statement or propagating irrelevant lines of thought about ideas peripheral to the main concept under investigation. The superior performance of the P/CP section and the Vignette sections (composed in a P/CP format as well) indicated that converting the T/F items into a P/CP format would (and thus has) increase(d) their validity.

4.1.4.3 Structural Validity

This category refers to the degree to which relationships between items conform to the theoretical view of the construct. The internal model for the construct specified *p*-value fluency to include proficiency across two sub-categories of *p*-value misinterpretations (single *p*-values, comparisons and predictions) as exhibited in both context-free items and when evaluating the research of others. This structure is reflected in the instrument as the test specifications address both facets of the research process and both sub-categories of misinterpretations (see the test blueprint for details).

The internal structure of the *model* suggested potential multidimensionality (two sub-categories of misinterpretations measured in two different settings), but factor analysis was necessary to investigate consistency with the internal structure suggested by the *data*. Three EFA models were checked (1, 2, and 3 factors), along with two CFA (both 2 factor) and two Bifactor models. Assessment of dimensionality was determined by the following indices: -2 log likelihood, AIC, and BIC (Table 4.32).

Table 4.32 – Diagnostics for Dimensionality Analysis[39]

Combined Sample - All Items			
Model	**-2LogLikelihood**	**AIC**	**BIC**
Bifactor (by item type)	5863.76	6079.76	6402.73
Bifactor (by misinterpretation type)	5849.53	6065.53	6388.5
EFA1(2PL)	5953.9	6097.9	6313.21
CFA2/2PL (by item type)	6007.63	6151.63	6366.94
CFA2/2PL (by misinterpretation type)	6032.4	6176.4	6391.71
EFA3/2PL	5819.86	6101.86	6523.51
EFA2/2PL	5881.59	6095.59	6415.56

Examination of the factor loadings for the 3-factor EFA model did not reveal a pattern that could be theoretically supported. Of the remaining two candidates, the 1-factor model and the Bifactor (by misinterpretation type) model, both support the general idea of a single underlying factor. Without compelling evidence against this, the research will maintain the position of unidimensionality at this time, but will keep an open mind as to the reevaluation of this assumption moving forward. Under a unidimensional assumption, the diagnostics for four model types (1PL, 2PL, 3PL, and Rasch) were analyzed. Results follow in Table 4.33 and indicate that the 2PL model is the clear favorite.

Table 4.33 – Diagnostics for Unidimensional Model Selection[40]

Combined Sample - All Items						
Model	**S.E.**	**χ2 ItemLevel Diag.**	**LID**	**-2LogLikelihood**	**AIC**	**BIC**
Rasch	0 items	8 items	11 items	6178.21	6250.21	6357.87
1PL	0 items	4 items	3 items	6139.37	6213.37	6324.02
2PL	3 items	2 items	1 item	5954.15	6098.15	6313.47
3PL	6 items	4 items	2 items	5976.32	6192.32	6515.29

[39] Full model output can be found in Appendix Y.
[40] This model output can also be found in Appendix Y.

4.1.4.4 Generalizability

This category refers to the degree to which test score properties and interpretations generalize to and across population groups, settings, and tasks. The target population for this instrument is future researchers: ideally, PhD candidates (outside the field of Statistics), who have completed their methodological training. It was the intention of the sampling process to capture a broad range of student characteristics to be inclusive of all genders, races, and programs of study. Demographic information, by Race/Ethnicity, Sex/Gender, and Status can be seen in Figure 4.26. As to generalizability, there appear to be no issues with regard to Sex/Gender or Status; however, the distribution of the Race/Ethnicity variable is concerning – particularly the lack of African American participation.

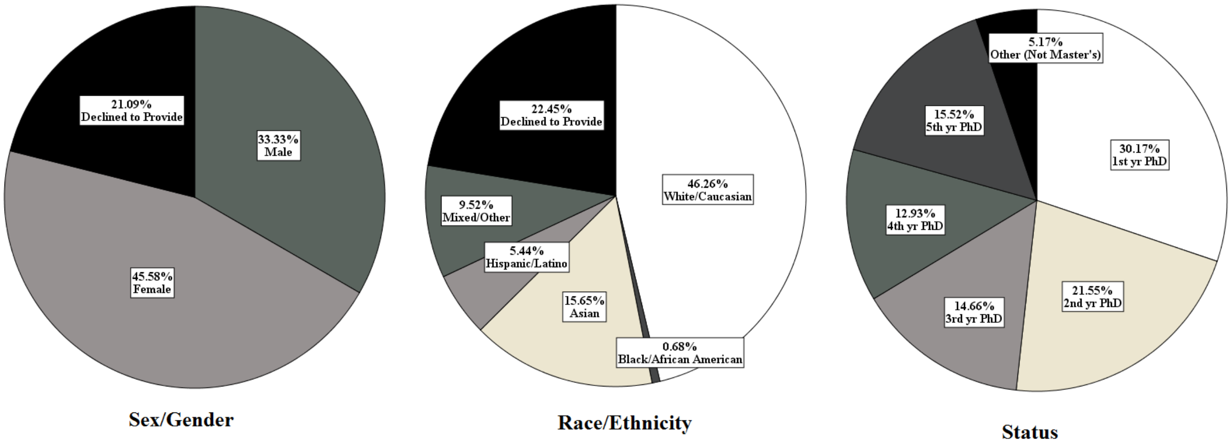


Figure 4.26. Demographic Characteristics of Full Dataset

The sample was meant to be nationally representative; however, with only 16 institutions identified by name, it is unlikely that this claim is supported. Furthermore, volunteer samples always suffer to some degree where generalizability is concerned in comparison to probability samples and thus this must be taken into consideration.

The demographic information was compiled to facilitate performance comparisons by subgroup and to allow Differential Item Function (DIF) analysis to be performed when possible[41]. This investigation concluded that DIF is not a concern when comparing by Sex/Gender, Race/Ethnicity, Status, or Institution based on the scant number of identified items and respective severity level classifications. Lack of DIF indicates that this instrument is generalizable in administration – i.e., exhibited item fairness across subgroups.

A key component of generalizability is the overall reliability of the instrument as measured by the Coefficient Alpha and related estimates. These estimates are affected both by length of instrument and size/nature of the sample; thus, these statistics were examined with an eye for modifications therein. Based on the results of the EFA, a unidimensional structure was assumed and thus there was no reason to compute reliability measures for sub-scales at this time. Table 4.34 shows that the reliability index was reasonable for this instrument (*Coefficient alpha* = .7732), as was the standard error (*SEM* = 2.7575).

Table 4.34 – Reliability Diagnostics

Test Level Statistics	
# of Items = 36	Mean = 20.0068
# of Examinees = 147	Standard Deviation = 5.7706
Minimum = 4.0	IQR = 7.0
Maximum = 32.0	Skewness = 0.0235
Median = 19.0	Kurtosis = -.4095

Reliability Analysis

Method	Estimate	95% CI	SEM
Guttman's L2	0.791	(0.7395, 0.8367)	2.6469
Coefficient Alpha	0.7732	(0.7173, 0.8227)	2.7575
Feldt-Gilmer	0.7835	(0.7301, 0.8308)	2.694
Feldt-Brennan	0.779	(0.7246, 0.8273)	2.7218
Raju's Beta	0.7732	(0.7173, 0.8227)	2.7575

[41] Not all cluster sizes were large enough to support this analysis. In those cases, a conclusion of *No DIF* cannot be reached. The determination that DIF is not a concern for this instrument was based on those comparisons that were able to be analyzed. Sparse categories would need to be investigated upon the receipt of additional data. All jMetrik output can be seen in Appendix Z.

Secondary analysis was performed using Item-deleted reliability measures. Ideally, these values should be lower than the overall reliability. Measures that are higher than the overall coefficient indicate that the instrument would function better if that item were removed. Items that demonstrated cause for concern were noted in Table 4.35. There were a surprisingly high number of flagged items (nine) in the P/CP section, but only one in the Vignette section. It is worth nothing; however, that despite the large number of items, the marginal increase in reliability is quite small for the majority of these items and removal is not likely to make an appreciable difference to the reliability of the instrument unless removed *en masse*.

Table 4.35 – Item Deleted Reliability (All Items)

Reliability if Item Deleted			
Item	**Alpha**	**Item**	**Alpha**
M12	0.7867	V3A1	0.7608
M3	0.7671	V3A3	0.7674
M1	0.7710	V3A9	0.7684
M13	0.7759	V3A17	0.7599
M18	0.7771	V3B4	0.7589
M9	0.7723	V3B12	0.7681
M4	0.7674	V3B15	0.7787
M16	0.7680	V3B16	0.7608
M14	0.7749	V5A7	0.7564
M6	0.7749	V5A10	0.7621
M11	0.7796	V5A11	0.7649
M7	0.7678	V5A14	0.7638
M5	0.7768	V5A18	0.7648
M2	0.7784	V5B2	0.7610
M10	0.7748	V5B5	0.7637
M15	0.7673	V5B6	0.7596
M8	0.7694	V5B8	0.7568
M17	0.7689	V5B13	0.7535

4.1.4.5 External Validity

This category refers to how the construct is expected to relate to other constructs and variables. One such approach is looking at measures of convergent and discriminant validity – i.e., comparing scores on the proposed instrument to scores on instruments known to be measuring the same/different constructs. That type of investigation was not undertaken in this research due to the nature of the recruitment and incentivization; however, student status (year in program) and coursework have been serving as a proxy thus far and will continue to do so here.

Table 4.36 – Linear Regression Model of Score as a Function of Status and Coursework

			Model Summary		
R	**R-square**	**Adjusted R-Square**	**Std. Error of the Estimate**		
0.403	0.162	0.147	0.141211		
			ANOVA		
	Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
Regression	0.413	2	0.207	10.365	<.001
Residual	2.134	107	0.02		
Total	2.547	109			
			Coefficients		
Model	**Unstandardized Coeff.**	**Std. Error**	**Standardized Coeff.**	**t**	**Sig.**
Constant	0.442	0.36		12.293	<.001
Status	0.18	0.009	0.176	1.989	0.49
Experience	0.41	0.01	0.363	4.099	<.001

The external model (Figure 3.2) situated statistical proficiency as a necessary component of PhD scholarship; therefore, it stands to reason that there is a continuum of proficiency that is navigated as one progresses through a plan of study. In other words, performance is likely to improve over time; thus, upper-division students (4th or 5th year) should outperform incoming students. Additionally, due to the varied nature of program requirements amongst PhD fields, there is a reasonable expectation that groups with differing levels of statistical preparation would perform disproportionately on this instrument – specifically that student performance would be directly related to experience (i.e., more coursework in statistical methods = higher scores). To investigate the hypothesized relationship between Status and Coursework with performance, a linear regression model was used. From this, we see that both indicators are statistically significant predictors of total score, projecting an average performance bump of ~2% per year in school and ~4% for each additional course completed (See Table 4.36). This result is consistent with expectation and thus serves to endorse the idea of external validity for this instrument.

All validity evidence thus far has been gathered on data from complete responses. There is reason to believe that data lost from those who abandoned the assessment in progress might not be missing completely at random. If patterns were to emerge that point to discrimination against students of identified sub-groups, this would affect the external validity; however, that analysis is impracticable for the data in hand and will not be conducted.

4.1.4.6 Consequential Validity

This category refers to the value implications of score interpretation as well as the actual and potential consequences of test use, especially regarding fairness. In the technical review of the items, the language was examined for use of offensive or stereotypic language. The items were written in a manner that rendered them devoid of ethnically-identifying pseudonyms and gender-specific pronouns, or at the very least, distributed them fairly. As mentioned previously, DIF analysis was performed on a variety of proposed subgroupings (Sex, Race, Status, and Institution) and this analysis raised no concerns.

The purpose of this instrument is to measure graduate students' proficiency in aspects of a statistical disposition most pertinent to their careers; e.g., communicating statistical results to others and contradicting claims made without proper statistical foundation. The results of the assessment will discriminate between those students who are well-prepared to be fully-functioning members of the research community and those who need more preparation in this aspect of their plan of study. The use of this instrument as designed may result in stigmatization of certain groups (i.e., those unlikely to score well); however, further analysis is required in order to determine whether this is a source of invalidity (e.g. lack of access) or an inevitable consequence that accompanies any sort of high-stakes testing.

The intended application of this instrument is to identify students who possess a lack of statistical and research knowledge and to identify programs for which there is a shortage of adequate statistical preparation. In doing so, the information can be used to design remediation programs for individuals and to institute policy change at the department level in terms of plan of study requirements. It remains to be seen what the consequences of these actions might entail (individual student attrition, burden on department resources, etc.) and further analysis is required in this aspect of validity but is presently inestimable with only the pilot/field test data at hand. Sparse data and small sample size make calculation of person-fit estimates impossible at the present time and thus will not be part of this dissertation.   This limitation is noted and discussed further in Chapter 5 with the expectation that this will be addressed in a future iteration of data collection.

4.1.4.7 Summary of Validation Plan Results

The validation plan for this instrument was designed to be comprehensive in addressing the six components of validity proposed by Messick: Content, Substantive, Structural, Generalizability, External, and Consequential. Analysis conducted within each component was undertaken with the aim of improving individual item quality and the performance (and thus implications) of the instrument as a whole. Taking a holistic view of the validation process, the instrument performed satisfactorily at this stage of development when considering the size and nature of the sample. A summary of validation results by category is displayed in Table 4.37. An action plan was devised to address those aspects of the validation plan that were in need of improvement. Not all of the proposed improvements were feasible with the present data, but for those that were, the modifications were implemented as indicated.

Table 4.37 – Instrument Validation Plan Summary

Category	**Indicator**	**Validation Status**	**Action Plan**
Content	Content Relevance	Satisfactory	
	Content Representativeness	Satisfactory	
	Technical Quality	Room for Improvement	**Item Revision and/or Removal**
Substantive	Theoretical Framework	Satisfactory	
	Construct Map/Internal Model	Room for Improvement	**Category/Process Level Revision**
	Engagement w/ Theoretical Framework	Satisfactory	
Structural	Dimensionality	Inconclusive	**Repeat analysis with amended instrument**
	Model Selection	Satisfactory	
Generalizability	Representativeness of Sample	Room for Improvement	New round of data collection with probability sampling
	DIF	Satisfactory	
	Global Reliability	Room for Improvement	**Repeat analysis with amended instrument**
	Item-Deleted Reliability	Room for Improvement	**Item Removal**
External	Convergent Validity	Room for Improvement	Repeat analysis with actual measure and not proxy
Consequential	Item Language	Satisfactory	
	DIF	Satisfactory	
	Potential Stigmatization	Inestimable	Explore when person-fit data is available (new sample)
	Consequences of Scoring Decisions	Inestimable	Explore when person-fit data is available (new sample)

*Note: Only items in **red** were conducted in this dissertation. The other action items are reserved for a future iteration of the research.*

Up to this point in time, validation efforts were undertaken under the assumption that the instrument was being validated on the whole, i.e., that all items were broadly measuring a single underlying construct, p-value fluency, albeit at various process levels (with/out context). Moving forward to implementation of the action plan required consideration as to whether independent validation of each sub-set of items was the more appropriate tact. As the Vignette section showed superior performance to the P/CP section, those items were analyzed first.

Improvements to Content Validity were indicated for the *technical quality* subcategory. The revision or removal of items with unsatisfactory levels of discrimination was the recommended course of action. When considering the Vignette section only (Table 4.38), we see that this action is unnecessary as there are zero items for which the item discrimination is unsatisfactory. We also notice that all items have appropriate item difficulty measures as well.

Table 4.38 – Classical Item Analysis Results, Vignette Section

Item	**Difficulty**	**Std. Dev.**	**Discrimination**
V3A1	0.6054	0.4904	0.5796
V3A3	0.3496	0.4904	0.3289
V3A9	0.3673	0.4837	0.3597
V3A17	0.6871	0.4653	0.6105
V3B4	0.6054	0.4904	0.5796
V3B12	0.7279	0.4466	0.5863
V3B15	0.3333	0.473	0.0913
V3B16	0.4218	0.4955	0.5755
V5A7	0.4694	0.5008	0.6036
V5A10	0.449	0.4991	0.5131
V5A11	0.4082	0.4932	0.5317
V5A14	0.6803	0.468	0.6050
V5A18	0.4558	0.4997	0.4209
V5B2	0.4694	0.5008	0.5561
V5B5	0.5306	0.5008	0.5376
V5B6	0.4898	0.5016	0.6141
V5B8	0.6054	0.4904	0.7147
V5B13	0.6667	0.473	0.8395

Improvements to Generalizability were indicated for the *global reliability* and *item-deleted reliability* sub-categories. The recommended course of action was the removal of items with high item-deleted reliabilities and recalculation of the global measure. An immediate, but unexpected, result was the improvement in reliability (as measured by *Coefficient Alpha*) that accompanied consideration of the vignettes as a stand-alone measure. That the global measure improved (.8298 vs. .7732) was particularly note-worthy because all things being equal, reliability measures favor longer instruments. Furthermore, where previously there were quite a few items identified as lowering the overall reliability, now there are only two such items (see Table 4.39)[42].

Table 4.39 – Item Deleted Reliability, Vignette Section

Reliability if Item Deleted			
Item	**Alpha**	**Item**	**Alpha**
V3A1	0.8195	V5A10	0.8221
V3A3	0.8300	V5A11	0.8215
V3A9	0.8287	V5A14	0.8193
V3A17	0.8192	V5A18	0.8261
V3B4	0.8195	V5B2	0.8202
V3B12	0.8207	V5B5	0.8210
V3B15	0.8389	V5B6	0.8176
V3B16	0.8195	V5B8	0.8137
V5A7	0.8181	V5B13	0.8095

Improvements to Structural Validity were investigated in the *dimensionality* analysis, but not really achieved. The previous investigation (with the full set of items) was inconclusive in determining the underlying latent structure. There was some evidence to suggest unidimensionality, but this was far from compelling. Analogous comparison of model diagnostics on the Vignettes only subset was similarly inconsistent (see brief results in Table 4.40 and full details in Appendix AA).

[42] Although convention would dictate that these two items be removed (V3A3 & V3B15), the marginal increase in *Coefficient Alpha* (.8396 vs. .8298) was not sufficiently large to justify the resultant failure to assess the theoretical framework in its entirety (i.e. not all misinterpretations would be tested).

Table 4.40 – Diagnostics for Dimensionality, Vignette Section

Combined Sample - Vignette Items			
Model	**-2LogLikelihood**	**AIC**	**BIC**
Bifactor (by item type)	N/A	N/A	N/A
Bifactor (by misinterpretation type)	2770.9	2878.9	3040.39
EFA1(2PL)	2811.63	2883.63	2991.29
CFA2/2PL (by item type)	2837.61	2909.61	3017.27
CFA2/2PL (by misinterpretation type)	N/A	N/A	N/A
EFA3/2PL	2746.18	2890.18	3105.49
EFA2/2PL	2771.72	2879.72	3041.2

A parallel investigation was undertaken to determine whether the P/CP items demonstrated sufficient performance as a stand-alone measure. Improvements to Content Validity were achieved in the *technical quality* subcategory. The item analysis (Table 4.41) revealed several items with unsatisfactory (negative or low) levels of discrimination.

Table 4.41 – Classical Item Analysis Results, P/CP Section

Item	Difficulty	Std. Dev.	Discrimination
M12	0.5238	0.5011	-0.2181
M3	0.5374	0.5003	0.4206
M1	0.483	0.5014	0.4118
M13	0.5714	0.4966	0.1513
M18	0.5102	0.5016	-0.0417
M9	0.449	0.4991	0.4400
M4	0.7415	0.4393	0.4813
M16	0.6395	0.4818	0.4087
M14	0.8571	0.3511	0.0267
M6	0.5102	0.5016	0.1754
M11	0.7755	0.4187	0.0459
M7	0.7891	0.4093	0.4540
M5	0.4286	0.4966	0.2967
M2	0.4626	0.5003	-0.0315
M10	0.4422	0.4983	0.1169
M15	0.5238	0.5011	0.1677
M8	0.7687	0.4231	0.2569
M17	0.6259	0.4856	0.3290

The first action was the removal of the three items (M12, M18, and M2) with negative item discriminations. The item analysis was then repeated on the remaining subset of items and re-examined for potentially problematic items. The update results (Table 4.42) identified three items as having item discriminations that, while positive, were near to zero and indicative of poor performance. These items were flagged but not yet removed. Item difficulties were considered satisfactory for all items (although *M14*was noted as being potentially *too easy*).

Table 4.42 – Classical Item Analysis, P/CP Section, Revised

Item	Difficulty	Std. Dev.	Discrimination
M3	0.5374	0.5003	0.4875
M1	0.483	0.5014	0.3999
M13	0.5714	0.4966	0.2167
M9	0.449	0.4991	0.3889
M4	0.7415	0.4393	0.5791
M16	0.6395	0.4818	0.4821
M14	0.8571	0.3511	0.1332
M6	0.5102	0.5016	0.2574
M11	0.7755	0.4187	0.0933
M7	0.7891	0.4093	0.5464
M5	0.4286	0.4966	0.2782
M10	0.4422	0.4983	0.0604
M15	0.5238	0.5011	0.1170
M8	0.7687	0.4231	0.3191
M17	0.6259	0.4856	0.3633

Improvements to Generalizability were indicated for the *global reliability* and *item-deleted reliability* sub-categories. The recommended course of action was the removal of items with high item-deleted reliabilities and recalculation of the global measure. The overall reliability of the P/CP section was notably weaker than when considering the Vignette section or the instrument as a whole (*Coefficient Alpha* = .5032, .8298, .7732, respectively). The removal of the three items with negative discriminations did improve this measure somewhat (.6128), but not to an acceptable level. The item-deleted reliability scores (Table 4.43) identified an additional three items as candidates for removal (m10, m11, and m15) – uncoincidentally, the same three items identified as having questionably low discrimination. Removal of these three items resulted in a small increase in *Coefficient Alpha* (.6512), and further identified this as the stabilization point for this measure. Item removal was discontinued at this point as it was determined that the removal of additional items would have negligible effect.

Table 4.43 – Item Deleted Reliability, P/CP Section

Reliability if Item Deleted			
Item	**Alpha**	**Item**	**Alpha**
M3	0.5694	M11	0.6023
M1	0.5821	M7	0.6211
M13	0.6079	M5	0.5750
M9	0.5839	M10	0.5995
M4	0.5664	M15	0.6287
M16	0.5726	M8	0.6214
M14	0.6160	M17	0.5977
M6	0.6023		

Improvements to Structural Validity were suggested in the *dimensionality* sub-category, but not undertaken for this set of items. With only 12 of the 18 original misinterpretations remaining, the reduction in content coverage was sufficient to disqualify the P/CP section from functioning as a stand-alone assessment of the construct thereby rendering the determination of dimensionality meaningless at this time.

4.1.5 Summary of Research Question 1

The culmination of the pilot testing, field testing, and validation procedures was supposed to be the identification of a sufficiently valid and reliable subset of items that would serve as the (current) final[43] version of the instrument. To that end, a collection of 30 items (12 from the P/CP section and 18 from the odd Vignette section), hereafter referred to as *PV3*, will serve in that capacity. That conclusion was reached as a plausible compromise between two seemingly irreconcilable positions. Firstly, the belief that instrument validation is an iterative process that is never truly complete. Any infusion of data or item modification will dictate that all measures

[43] The *final* version of the instrument in terms of *this* research study. As discussed in Chapter 5, the instrument validation is a work in process and future data collection and instrument revision is planned.

(e.g. reliability, item discrimination, etc.) be updated accordingly. And on the other hand, the notion that validation is necessarily a means to an end: if one never *completes* validation, no instrument will exist to achieve the desired ends (i.e., drawing conclusions from scores).
At some point in the process, the models and indices will (must?) stabilize to a point of utility. Indubitably, the *KPVMI* assessment has not yet achieved that level of technical quality with the *PV3* iteration; however, incompleteness in and of itself does not nullify the validation efforts therein.

All validation efforts to date have been in pursuit of answering the first research question, *Can a sufficiently reliable and valid measure of p-value misinterpretations (in a research context) be constructed*? On the premise that instrument validation is never *over,* a better question would have been, *Is there enough validity evidence to date to support the use of this instrument?* In response, it is reasonable to declare that the initial validation efforts were promising and use of the instrument would be justified in its current state, keeping in mind that despite the modifications already implemented, there is still potential for improvement. The PV3 iteration of the instrument, when considered on the whole, is a moderately reliable measure (*Alpha* = .8030) of p-value fluency as assessed across 18 misinterpretations (the entirety of the theoretical framework) and 2 process levels; when considered in part, it contains an independently validated sub-measure of *p*-value fluency *in context* as assessed across all 18 misinterpretations (*Alpha* = .8298).

That being said, a major unresolved conflict remained that presented an obstacle in addressing the second research question, namely the specification of the dimensionality of the latent space. The existing test blueprint was predicated on the notion that the main construct, *p*-value fluency, potentially has sub-factors based on item type (with/out context) or misinterpretation type (single *p*-values vs. comparisons/predictions). Without establishing the unidimensionality of the construct, the consequences of removal of an entire section of items is inestimable. If the construct is truly unidimensional, then any sufficiently reliable combination of items will suffice (spanning item types and misinterpretation types) and thus there would be no reason why the Vignette section could not stand alone (selected on the basis of its superior performance to both the P/CP subset and the PV3 30-item version). On the other hand, if there exist factors by item and/or misinterpretation type, then it would be inappropriate to posit the Vignette section as a suitable version of the instrument and not merely a properly validated sub-section.

The intention of this section of the paper was to conclude the validation plan discussion with a presentation of the updated test blueprint (consistent with the PV3 version of the instrument) meant to reflect the confirmed latent structure of the p-value fluency construct and use that to answer the second research question and as the basis for future administrations of the assessment. As will be discussed further in Chapter 5, the inability to resolve this conflict is a limitation of the work that will be addressed with future data collection. As this issue is yet unresolved, multiple candidate blueprints are presented. The *PV3* blueprint presents the *actual* distribution of the existing 30 items (Table 4.44). Also presented, in the subsequent table, are two proposed blueprints: *KPVMI*(P) and *KPVMI*(F) for *partial* and *full*, respectively (Table 4.45). Both of the proposed versions, along with that for PV3, assume one underlying content dimension (in contrast to the PV2 version which separated the *single p-value* misinterpretations from that of *comparisons and predictions*), but differ in the assignment of items to process levels. The presentation of two blueprint options is meant to facilitate future investigation into dimensionality of the latent space while also introducing the opportunity for the existence of both full-length and abbreviated versions of the assessment down the road.

Table 4.44 – PV3 Test Blueprint

Test Blueprint - PV2 Version		
	Process Level	
Item Content	**Context-Free**	**In Context**
	PCP	Odd Vig
Misinterpretations of single P-values	14	14
MI1	1	1
MI2	1	1
MI3	1	1
MI4	1	1
MI5	1	1
MI6	1	1
MI7	1	1
MI8	1	1
MI9	1	1
MI10	1	1
MI11	1	1
MI12	1	1
MI13	1	1
MI14	1	1
Misinterpretations of P-value comparisons and predictions	4	4
MI15	1	1
MI16	1	1
MI17	1	1
MI18	1	1
2 sub-dimensions tested across 2 process levels.		

Test Blueprint - PV3 Version		
	Process Level	
Item Content	**Context-Free**	**In Context**
	PCP	Odd Vig
Misinterpretations of P-values	12	18
MI1	1	1
MI2		1
MI3	1	1
MI4	1	1
MI5	1	1
MI6	1	1
MI7	1	1
MI8	1	1
MI9	1	1
MI10		1
MI11		1
MI12		1
MI13	1	1
MI14	1	1
MI15		1
MI16	1	1
MI17	1	1
MI18		1
1 dimension tested across 2 process levels.		

Table 4.45 – Proposed Test Blueprints for *KPVMI(P)* and *KPVMI(F)*

Updated Test Blueprint – *KPVMI(P)* Version		**Updated Test Blueprint – *KPVMI(F)* Version**		
	Process Level		**Process Level**	
Item Content	**In Context**	**Item Content**	**Context-Free**	**In Context**
	Odd Vig		PCP	Odd Vig
Misinterpretations of P-values	18	Misinterpretations of P-values	18	18
MI1	1	MI1	1	1
MI2	1	MI2	1	1
MI3	1	MI3	1	1
MI4	1	MI4	1	1
MI5	1	MI5	1	1
MI6	1	MI6	1	1
MI7	1	MI7	1	1
MI8	1	MI8	1	1
MI9	1	MI9	1	1
MI10	1	MI10	1	1
MI11	1	MI11	1	1
MI12	1	MI12	1	1
MI13	1	MI13	1	1
MI14	1	MI14	1	1
MI15	1	MI15	1	1
MI16	1	MI16	1	1
MI17	1	MI17	1	1
MI18	1	MI18	1	1
1 dimension tested across 1 process level.		1 dimension tested across 2 process levels.		

For now, for the purposes of addressing the second research question, the *KPVMI(P)* version was adopted and all analysis was conducted on the Vignette subsection of items only. This decision ensured that any implications drawn about the *p*-value fluency of doctoral students nationally was based on the most reliable subset of available items, but required that the language of any such conclusions reflect the limited power of the abbreviated assessment (i.e., *p*-value fluency was only addressed *in context*, which may or may not represent the *easier* of potentially two process levels).

4.2 Research Question 2

The second research question: *What do the results of the KPVMI-1 administration tell us about the current level of p-value fluency among doctoral students nationally?* was addressed via analysis of a subset of the field test data (n = 147) with respect to performance on the subset of items considered sufficiently validated as developed in Phases I-III. In this section of the paper, as instrument validation remains a work in progress, we endeavor to answer that question in the limited capacity that the size and nature of the sample permit and using an abbreviated version of the instrument, the aforementioned *KPVMI(P),* hereafter referred to as the *KPVMI-1* to denote results of the *first* administration.

Appropriate descriptive statistics for overall performance, and by sub-category of misinterpretations where factor analysis indicated such a calculation was warranted, were performed. Secondary analysis compared performance figures by subgroup (e.g. by gender, by race, by program of study, by level of preparation, etc.) and made inferences therein. Specifically, this phase of the research investigated the following sub research questions: *Are there performance differences by subgroup? Can these differences be attributed to differences in experience or preparation? What are the implications of these results in the context of the ASA's principles?*

4.2.1 Overall Performance

In general, the median score was 10/18 items correct, or slightly better than half (56%). Many respondents scored better than this, but the distribution is left-skewed due to a large number of 0 scores, which lowers the mean to 9.37 (52%). When selecting these 147 cases for inclusion in the dataset, it was not a requirement that the responses be complete. Sparse entries were permitted so long as they exceeded the threshold of 15 items answered, but that was 15 items of the original 36. For many of the incomplete responses, the attrition occurred prior to or midway through completion of the Vignette section (presented last on Qualtrics) and thus these 0 scores do not actually represent responses that are entirely incorrect but rather entirely blank (in most cases). In light of that, the distribution curve and summary statistics were recomputed with these blank cases removed to give a more accurate representation (only removed if *completely* blank, any partial attempts at the section were retained). The median did not change; however, the mean did increase slightly to 10.2 (57%). Distribution curves are presented in Figure 4.27 and summary statistics are given in Table 4.46.

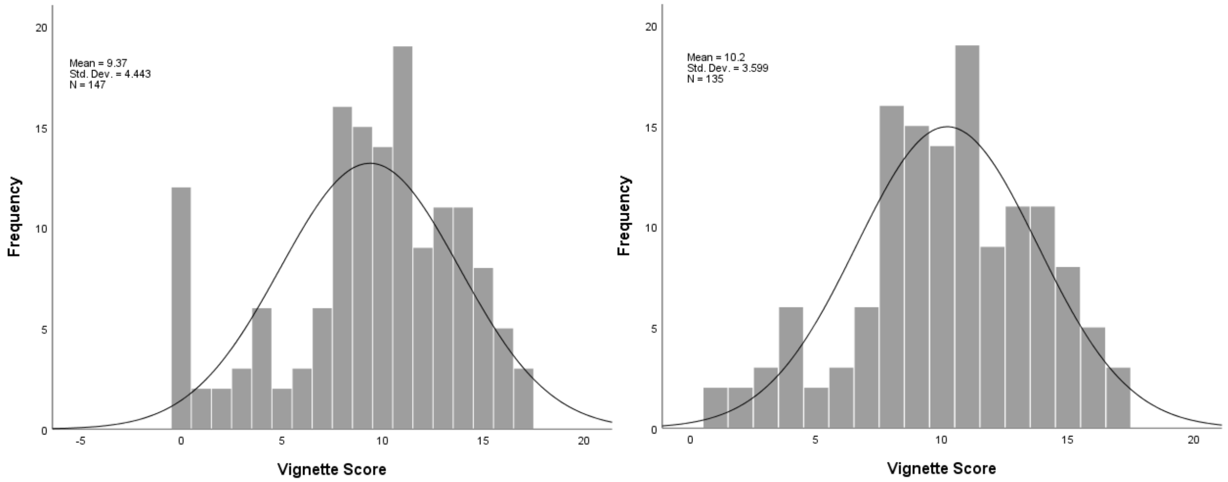


Figure 4.27. *KPVMI-1* Performance Distribution, National Sample

Table 4.46 – KPVMI-1 Performance Statistics, National Sample

Data		N	Minimum	Maximum	Mean	Std. Deviation
Full Sample	RV (%)	147	0	0.944	0.52041	0.246837
	RV (items)	147	0	17	9.37	4.443
Trimmed Sample	RV (%)	135	0.056	0.944	0.56667	0.19994
	RV (items)	135	1	17	10.2	3.599

Looking by item now, Figure 4.28[44] shows the ordered distribution of correct/incorrect responses, arranged in ascending order from #1-#18; Figure 4.29 shows the distribution arranged in descending order (correct responses) from *easiest* to *hardest*. In each figure, a reference line

[44] Note that this graph is scaled to 100% for each item. This does not indicate, and it is not the case, that the number of respondents was the same across all items. Responses ranged from n = 122 to n = 137.

is drawn at 50% for ease in identifying items for which fewer than half of the respondents answered correctly. What is particularly interesting about this distribution is the absence of identified clusters; in other words, the item difficulties seem to gradually trend downward like stair steps as opposed to bunching in clusters of *easy* and *hard* items with obvious breakpoints.

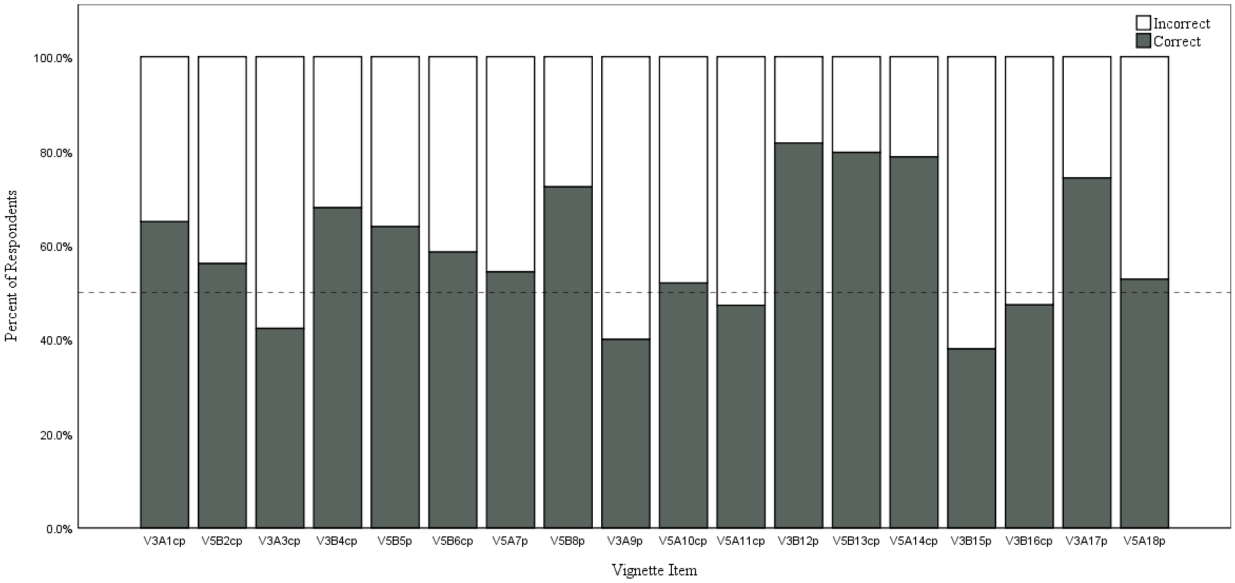


Figure 4.28. *KPVMI-1* Ordered Item Response Distribution, National Sample

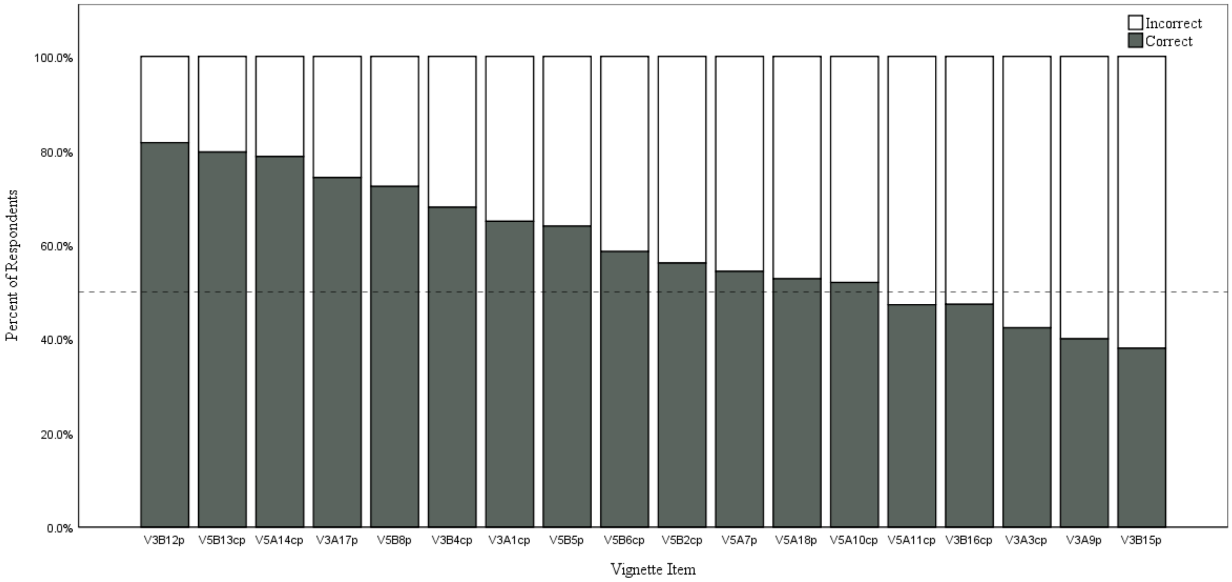


Figure 4.29. *KPVMI-1* Item Response Distribution by Difficulty, National Sample

Table 4.47 presents the item rankings paired with the misinterpretations. In this chart, smaller numbers indicate easier items (#1 indicates the item with the highest percentage of correct responses) and larger numbers indicate more difficult items.

Table 4.47 – Misinterpretation Difficulty Ranking, National Sample

Category	Rank	Misinterpretation
Common misinterpretations of single P values	7	1. The P value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave P = 0.01, the null hypothesis has only a 1% chance of being true; if instead it gave P = .40, the null hypothesis has a 40% chance of being true.
	10	2. The P value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the P value for the null hypothesis is 0.08, there is an 8% probability that chance alone produced the association.
	16	3. A significant test result (P ≤ 0.05) means that the test hypothesis is false or should be rejected.
	6	4. A nonsignificant test result (P > 0.05) means that the test hypothesis is true or should be accepted.
	8	5. A large P value is evidence in favor of the test hypothesis.
	9	6. A null hypothesis P value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.
	11	7. Statistical significance indicates a scientifically or substantively important relation has been detected.
	5	8. Lack of statistical significance indicates that the effect size is small.
	17	9. The P value is the chance of our data occurring if the test hypothesis is true; for example, P = 0.05 means that the observed association would occur only 5% of the time under the test hypothesis.
	13	10. If you reject the test hypothesis because P ≤ 0.05, the chance you are in error (the chance your "significant finding" is a false positive) is 5%
	14	11. P = 0.05 and P ≤ 0.05 mean the same thing.
	1	12. P values are properly reported as inequalities (e.g., report "P < 0.02" when P = 0.015 or report "P > 0.05" when P = 0.06 or P = 0.70)
	2	13. Statistical significance is a property of the phenomenon being studied, and thus statistical tests detect significance.
	3	14. One should always use two-sided P values.
Common misinterpretations of P value comparisons and predictions	18	15. When the same hypothesis is tested in different studies and none or a minority of the tests are statistically significant (all P > 0.05), the overall evidence supports the hypothesis.
	15	16. When the same hypothesis is tested in two different populations, and the resulting P values are on opposite sides of 0.05, the results are conflicting.
	4	17. When the same hypothesis is tested in two different populations and the same P values are obtained, the results are in agreement.
	12	18. If one observes a small P value, there is a good chance that the next study will produce a P value at least as small for the same hypothesis.

What is striking here is the apparent lack of consistency in the rankings. The rankings do not appear to support the notion that the *Misinterpretations of p-value comparisons and predictions* items are at a different process level than that of the *Misinterpretations of single p-values* items. Three of the items in the former category are fairly difficult (#18, #15, and #12), but the fourth item in this category is rather easy (#4). Similarly, the rankings do not appear to support the notion that similar misinterpretations would be consistently difficult across items. For instance, both misinterpretation #16 and #17 ask the respondent to compare *p*-values across a pair of studies (once indicating that the *p*-values are in agreement and once indicating that the *p*-values are conflicting, but essentially the same idea), yet item #16 is considerably more difficult than item #17 (ranking of 15 vs. ranking of 4). Likewise, misinterpretations #3 and #4 similarly ask the respondent to determine if the significance of the result (or lack thereof) determines the falsity (or truth) of the null hypothesis but item #3 is considerably more difficult than item #4 (ranking of 16 vs. ranking of 6).

4.2.2 Performance by Subgroup[45]

In this secondary analysis, the following sub- research questions are investigated: *Are there performance differences by subgroup? Can these differences be attributed to differences in experience or preparation?*

4.2.2.1 Performance Comparison by Sex

Male respondents outperformed female respondents slightly (*n* = 49, mean male score = 11.10; *n* = 67, mean female score = 10.33), but this difference was both statistically insignificant (*t* = 1.235, *p* = .220) and practically unimportant. The mean difference (.774) indicated that in

[45] Note that for all subgroup analyses, the data set was restricted to those respondents who not only answered the item but chose to identify themselves; hence calculations are based on a subset of the 147 respondents in the total sample and the size of the sample in question fluctuates between analyses. Also note that all Sex and Race distinctions are *self*-identified. Majors were clustered by the research team based on write-in responses.

general the sex discrepancy was less than a single item on average. The boxplot of the distributions (Figure 4.30) reveals that the female subgroup had outlying respondents on both ends of the performance spectrum, while the male subgroup was more consistent.

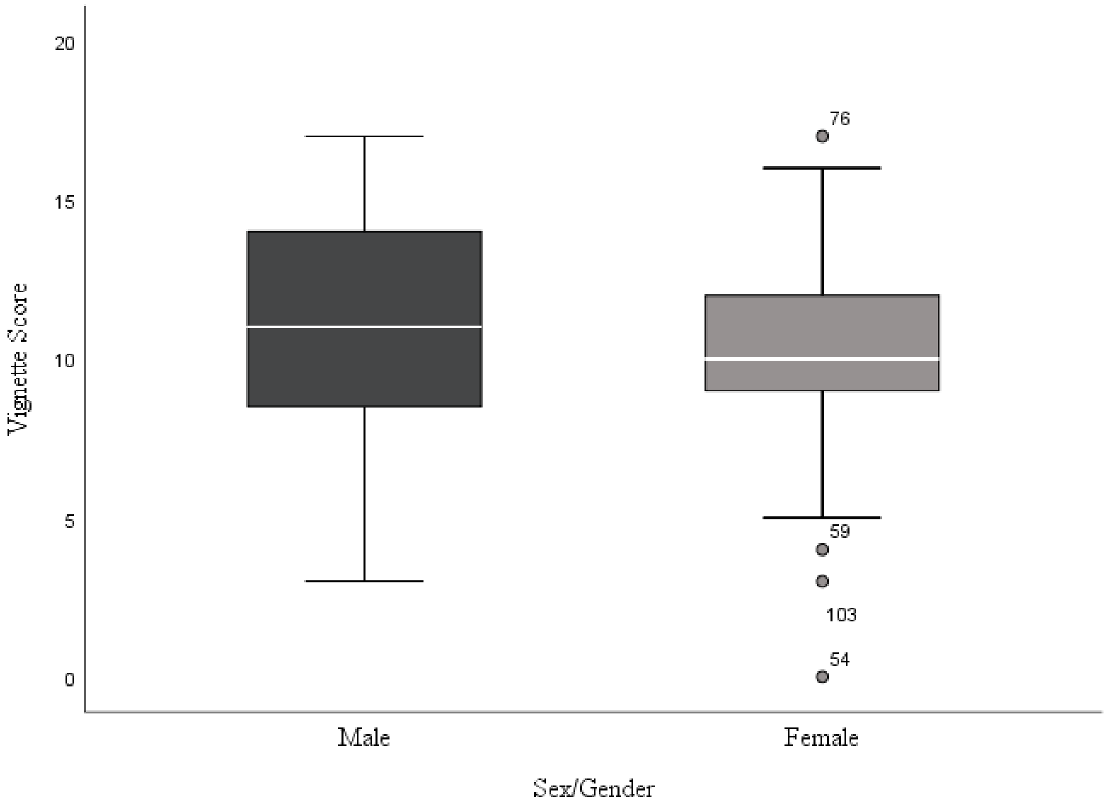


Figure 4.30. KPVMI-1 Performance by Sex/Gender

The male performance advantage did not warrant a follow-up investigation by virtue of its size; however, the trend was notable when considering that female respondents were equally experienced (average courses taken was 2.28 for female, 2.35 for male) but slightly more senior (average status was 2.74 years for female, 2.42 for male) suggesting that if anything, one might expect to see a slight female advantage. Score trends as distributed across levels of Status and Experience (i.e., courses taken) are displayed in Figure 4.31. We see that the male advantage remains consistent across the academic career (i.e., status), but that in terms of experience, female respondents perform as well or better when all students are inexperienced (i.e., less than 2 courses), but that male respondents tend to reap a larger marginal benefit from the additional coursework.

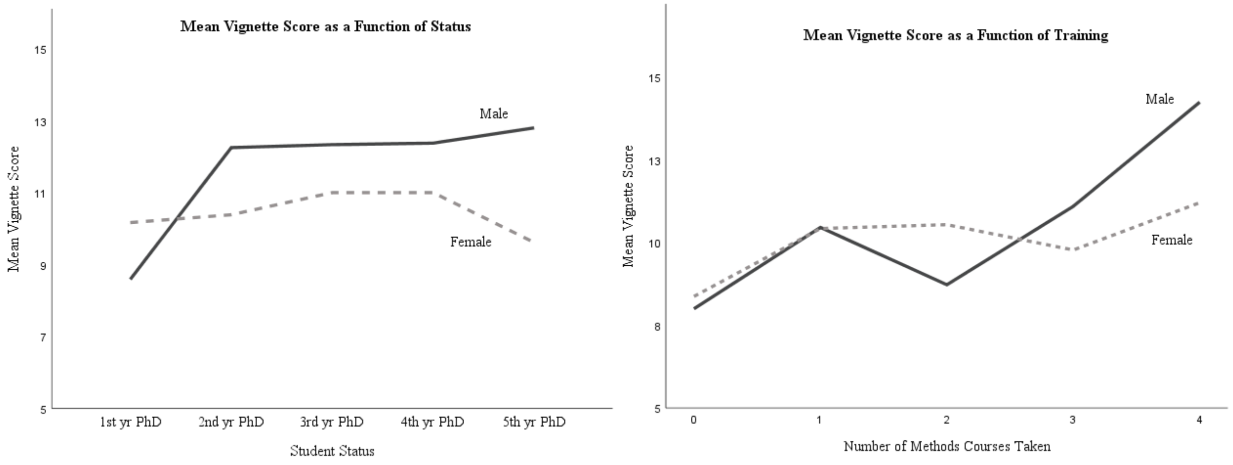


Figure 4.31. KPVMI-1 Performance as a Function of Status and Experience by Sex/Gender

4.2.2.2 Performance Comparison by Major

In decreasing order of performance, Business/Economics respondents scored the highest (*n* = 11, mean of 12.82), followed by Psychology/Sociology/Education respondents (*n* = 25, mean of 11.40), followed by Biology/Medicine respondents (*n* = 31, mean of 10.65), followed by Math/Physics/Engineering respondents (*n* = 27, mean = 10.11), and lastly, Languages/Communication/Anthropology respondents (*n* = 10, mean = 9.50). The 1-way ANOVA test for the group mean comparison was statistically insignificant at the .05 level (*F* = 2.181, *p* = .077); thus, no post-hoc testing was necessary. The boxplot of the distributions can be seen in Figure 4.32.

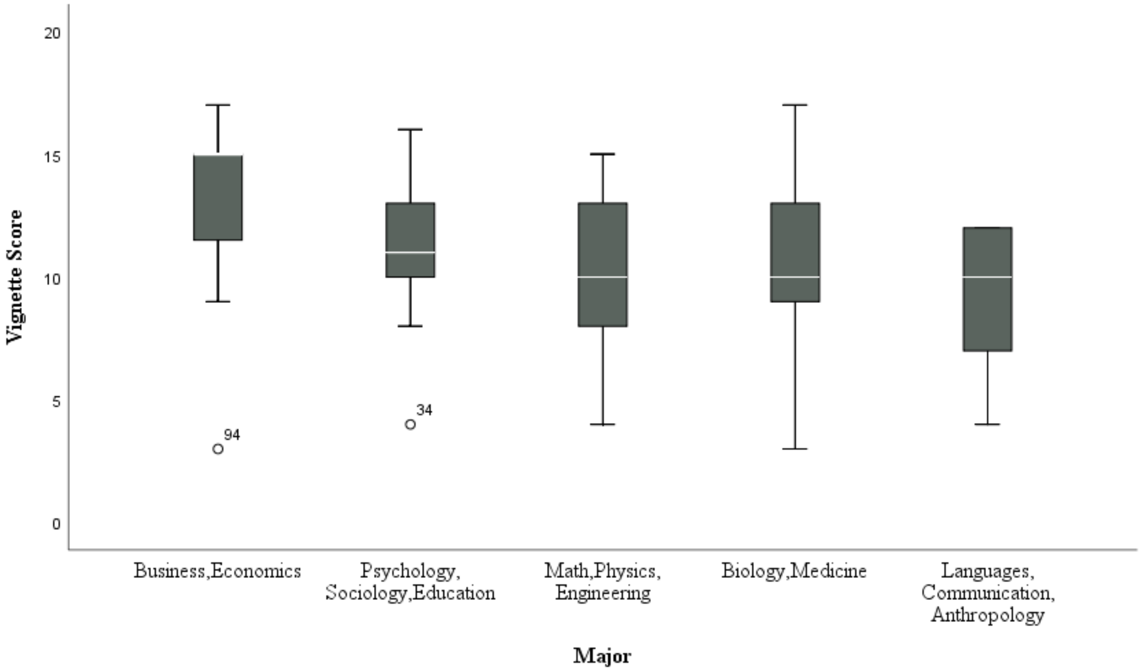


Figure 4.32. *KPVMI-1* Performance by Major

Here the group discrepancies are at times large enough to be considered *practically important* despite having not achieved statistical significance. For instance, the mean discrepancy between the Business/Economics respondents and the Languages/Communication/Anthropology respondents was 3.32 – nearly five times the mean discrepancy between the male and female respondents (which *was* statistically significant). A cursory glance at experience and status differences suggested an interesting relationship in that the degree programs with the more senior students (Biology/Medicine & Languages/Communication/Anthropology) report having taken the fewest courses. So, while it makes sense that Business/Economics respondents, having the most experience, would score higher than respondents in other majors, it doesn't necessarily make sense that these students would be the most junior. This suggests perhaps that course-taking patterns are not consistent across program types and that maybe the assumption that *more status = more courses* is not necessarily a valid assumption. Certainly, this finding seems to support the notion that experience (number of courses taken) is a more influential predictor of performance than status (year in program). See Figure 4.33 for details.

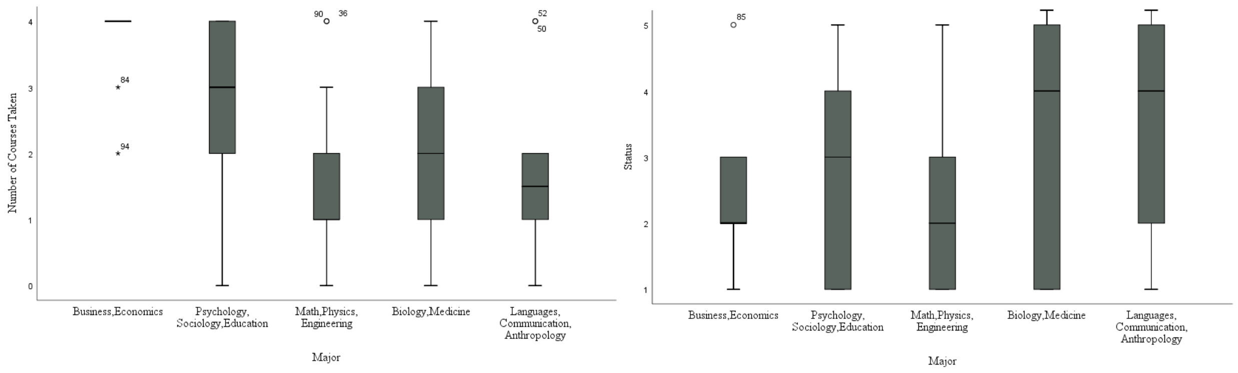


Figure 4.33. *KPVMI-1* Experience and Status by Major

Line graphs of score trends as distributed across levels of Status and Experience (i.e., courses taken) are not very elucidative (See Figure 4.34). It is difficult to make either the claim that increased training correlates with higher performance or the claim that seniority correlates with higher performance. A striking feature of these graphs is that there is no clear leader – for any given combination of program and either status/experience, the particular ranking of performance varies. This is an indication of an interaction effect between these variables, but that will not be investigated in this research due to sample size limitations.

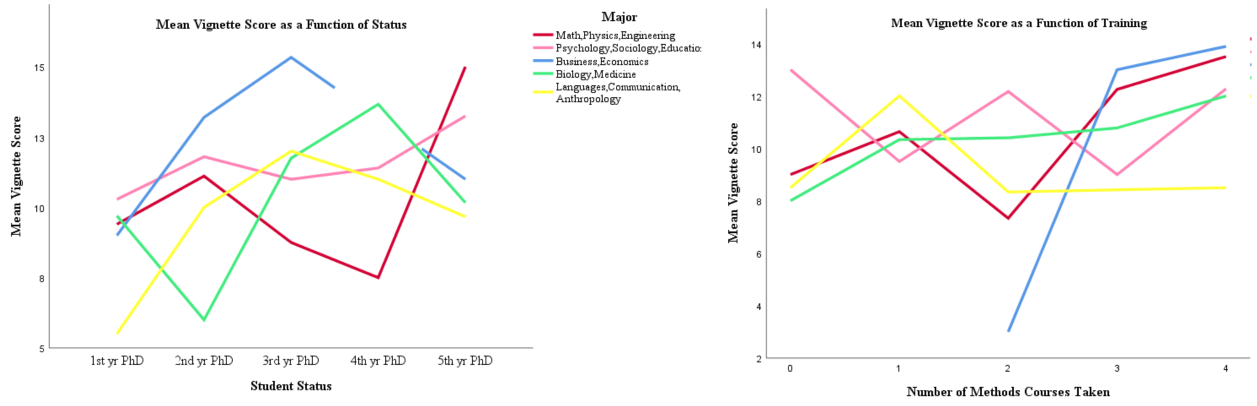


Figure 4.34. KPVMI-1 Performance as a Function of Status and Experience by Major

4.2.3 Performance by ASA Subscale

Having looked at overall scores and performance by subgroup, the last sub-investigation embedded in the second research question shifts the focus away from the scores themselves to the implications therein. Specifically, what can be inferred from these results in the context of the ASA's six principles for proper use and interpretation of the *p*-value[46]?

In drafting the language for its statement, the ASA was intentionally positive, in other words, the statement offers guidance on how *p*-values *should* be regarded and utilized, as opposed to a list of ways in which they are *not*. For a few select misinterpretations on the instrument's theoretical framework, the statement and the ASA principle are mere negations of each other and thus a 1-to-1 correspondence is rather straightforward (e.g., ASA principle #2: Misinterpretation #1). The functional relationship between the sets however was necessarily surjective (and not bijective) as there were many more misinterpretations than there were ASA principles. For those misinterpretations where the mapping was not patent, assignment was determined on the basis of which misinterpretations would interfere with the ability to understand and/or execute the stated guideline (see Table 4.48).

[46] The principles are listed in Section 2.3.3 of this document.

Table 4.48 – P-value Misinterpretations Cross-Referenced by ASA Principle

Category	ASA	Misinterpretation
Common misinterpretations of single P values	1	5. A large P value is evidence in favor of the test hypothesis. 9. The P value is the chance of our data occurring if the test hypothesis is true; for example, P = 0.05 means that the observed association would occur only 5% of the time under the test hypothesis. 10. If you reject the test hypothesis because P ≤ 0.05, the chance you are in error (the chance your "significant finding" is a false positive) is 5%
	2	1. The P value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave P = 0.01, the null hypothesis has only a 1% chance of being true; if instead it gave P = .40, the null hypothesis has a 40% chance of being true. 2. The P value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the P value for the null hypothesis is 0.08, there is an 8% probability that chance alone produced the association. 3. A significant test result (P ≤ 0.05) means that the test hypothesis is false or should be rejected. 4. A nonsignificant test result (P > 0.05) means that the test hypothesis is true or should be accepted.
	3	13. Statistical significance is a property of the phenomenon being studied, and thus statistical tests detect significance.
	4	11. P = 0.05 and P ≤ 0.05 mean the same thing. 12. P values are properly reported as inequalities (e.g., report "P < 0.02" when P = 0.015 or report "P > 0.05" when P = 0.06 or P = 0.70) 14. One should always use two-sided P values.
	5	6. A null hypothesis P value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated. 7. Statistical significance indicates a scientifically or substantively important relation has been detected. 8. Lack of statistical significance indicates that the effect size is small.
Common misinterpretations of P value comparisons and predictions	6	15. When the same hypothesis is tested in different studies and none or a minority of the tests are statistically significant (all P > 0.05), the overall evidence supports the hypothesis. 16. When the same hypothesis is tested in two different populations, and the resulting P values are on opposite sides of 0.05, the results are conflicting. 17. When the same hypothesis is tested in two different populations and the same P values are obtained, the results are in agreement. 18. If one observes a small P value, there is a good chance that the next study will produce a P value at least as small for the same hypothesis.

After assigning misinterpretations to principles, the corresponding test items were used to compute sub-scores. In practice, in future iterations of the *KPVMI* instrument, *all* items corresponding to a particular misinterpretation will be used in the sub-score calculation, but in the present analysis only the odd vignette items were considered[47]. Means and standard deviations for each of the six subscales are presented in Table 4.49. Note that the range for each is [0, 1] and a perfect score indicates that all items in that cluster were answered correctly.

Table 4.49 – KPVMI-1 ASA Subscale Performance, National Sample

		ASA1	**ASA2**	**ASA3**	**ASA4**	**ASA5**	**ASA6**
N	Valid	121	123	123	127	123	125
	Missing	26	24	24	20	24	22
	Mean	0.5234	0.5854	0.7967	0.6929	0.6125	0.54
	Standard Deviation	0.31572	0.29655	0.40406	0.2643	0.33435	0.24675

From this table, we observe that our respondents seem least likely to be susceptible to violations of Principle #3 and, in descending order, Principle #4, Principle #5, Principle #2, Principle #6, and finally, Principle #1.

Analogous to the subgroup comparisons of the previous section, a similar analysis was conducted whereby the ordered relationship of the subscales was investigated across three variables: Major (Figure 4.35), Status (Figure 4.36), and Experience (Figure 4.37). When analyzing these graphs, there are two important features to attend to. Firstly, overall trends in the subscale lines indicate the relationship between the categories in the classification scheme and the performance; rising (falling) lines indicate increases (decreases) in performance as when traversing the groups along the x-axis. Secondly, the degree of parallelism in the contours is an important indication of the dependency of that classification variable and the subscale performance. Parallel (or a reasonable approximation) contours would indicate an independence, i.e., that the ordered relationship between the subscales is consistent across categories. Intersecting trend lines would indicate the presence of an interaction effect.

[47] Despite the presence of a partial collection of validated PCP items, the decision to use only the vignette items was to keep consistent with the declaration made in Section 4.2 of this document.

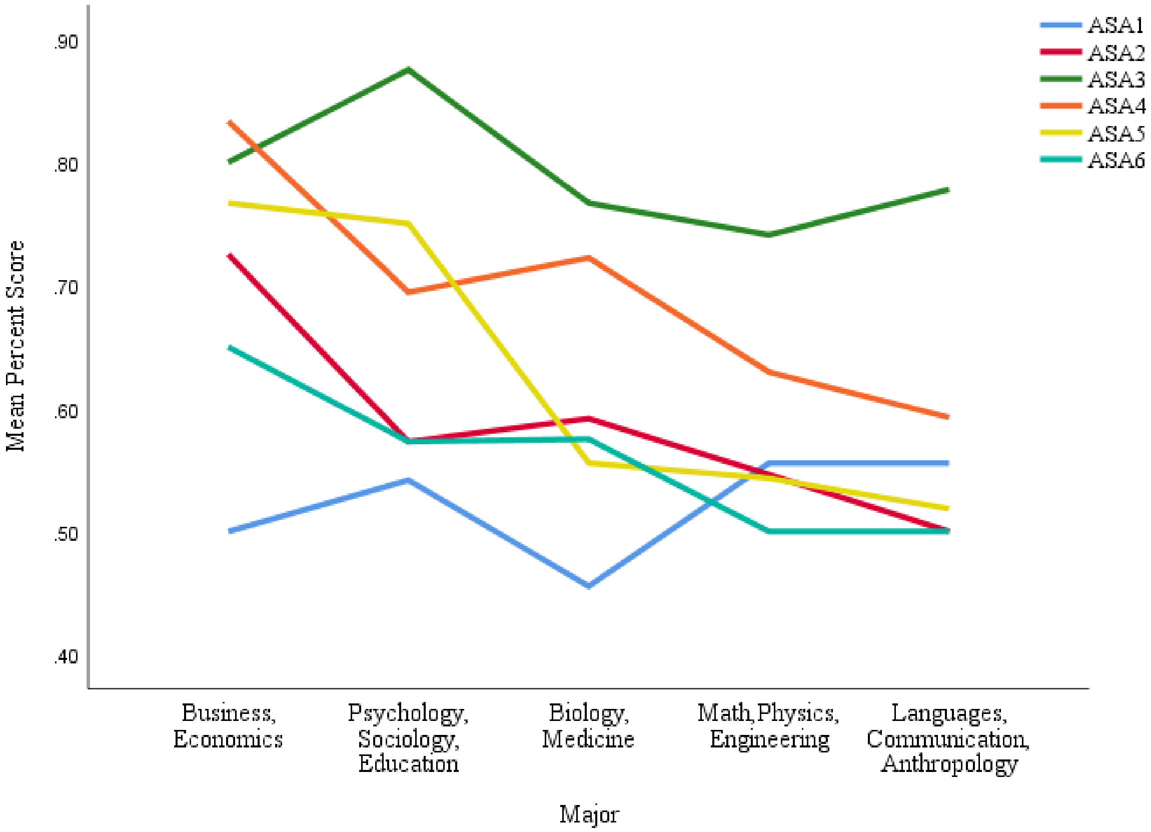


Figure 4.35. *KPVMI-1* ASA Subscores by Major, National Sample

Inspection of the subscores by Major suggests a (mostly) decreasing relationship: the categories, ordered based on overall performance, tend to indicate declining performance from left to right (i.e., Business/Economics, Psychology/Sociology/Education, Biology/Medicine, Math/Physics/Engineering, Languages/Communication/Anthropology). ASA1 is a notable exception to this observation. The contours, while not parallel, do provide some evidence that the subscale performance is somewhat independent of Major. ASA3 and ASA4 are (essentially) universally easier for respondents while ASA1 and ASA6 are universally more difficult. The unexpected precipitous drop-off of ASA5 does obscure this relationship a bit, but not enough to undermine the overarching sentiment.

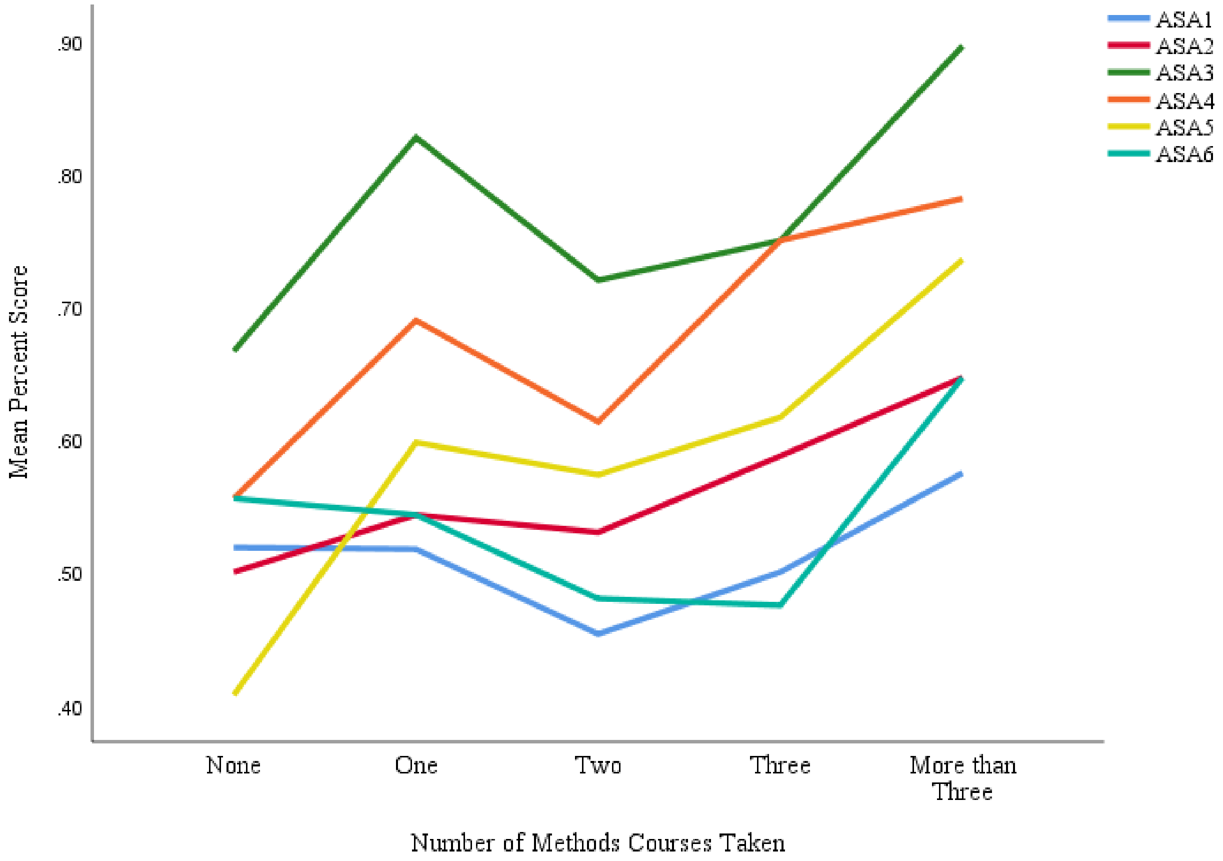


Figure 4.36. *KPVMI-1* ASA Subscores by Experience, National Sample

Inspection of the subscores by Experience suggests a (mostly) increasing relationship: performance tends to increase as the number of courses taken increases. It is important to appreciate that none of these trends are monotonic and in particular, ASA6 very nearly violates this assertion entirely. A plausible explanation for this pattern might be that brief downturns in the trend line might be illustrating the *little bit of knowledge is a dangerous thing* phenomenon whereby partial understanding can lead to misinterpretations in the short run that tend to be ameliorated with future study. The contours, again not parallel, do provide some evidence that the subscale performance is somewhat independent of Experience. Consistent with the previous results, ASA4 and ASA3 remain at the top while ASA1 and ASA6 remain the most elusive – across all levels of Experience. This is perhaps counterintuitive as one might assume that increased training might lead to a plateau point at which all misinterpretations have become extinct and thus all ASA subscales are equally understood.

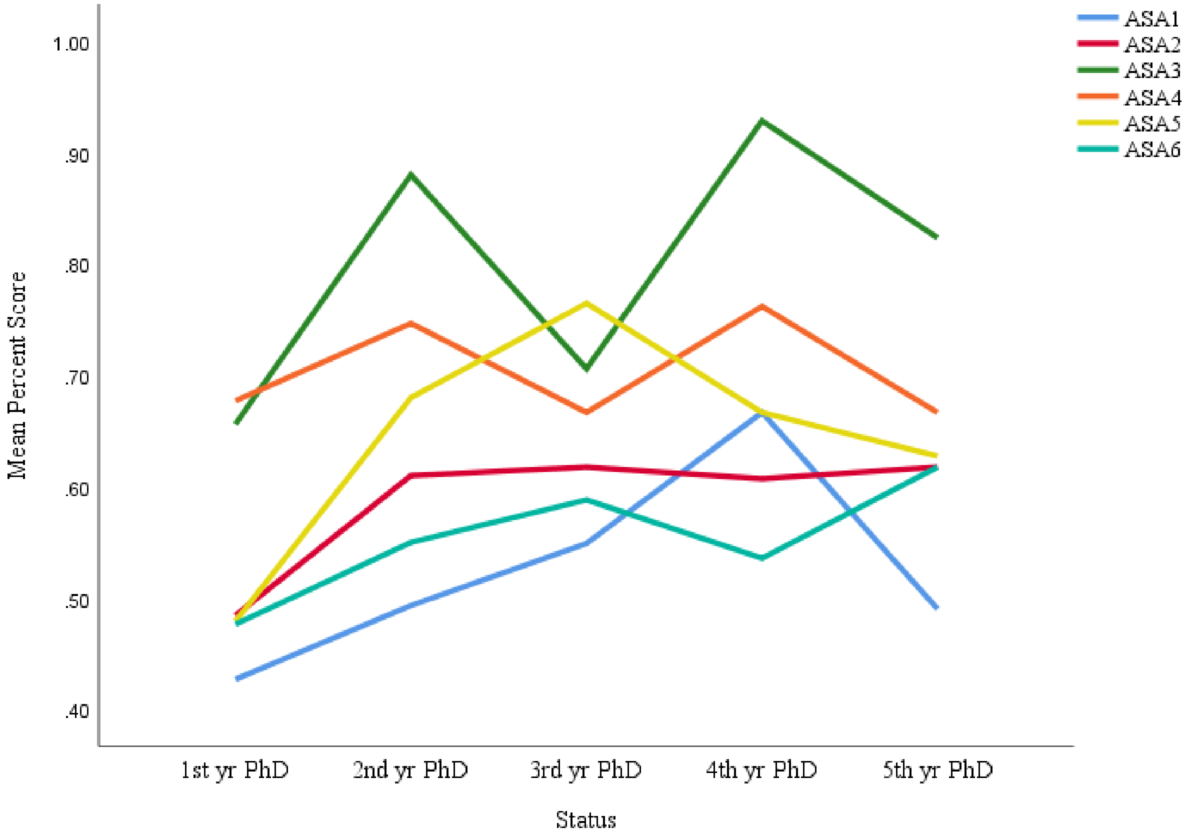


Figure 4.37. KPVMI-1 ASA Subscores by Status, National Sample

Inspection of the subscores by Status suggests a complicated relationship between performance and progression through one's program of study. Scores tend to ebb and flow over the five-year period, but not consistently so by subscale. The lack of parallelism in the contours strongly suggests an interaction between this classification method and subscale performance. Interestingly enough, ASA3 and ASA4 continue to take the top spots (basically) while ASA6 and ASA1 remain at the bottom of the spectrum – across all levels of Status. Also noteworthy is the peculiar behavior of ASA5 and the observation that four of the six subscores decline from the 4th year to the 5th. A plausible explanation for these findings might be the organic fluctuations in degree completion expectations by institution and Major. The natural assumption is that more senior students are likely more knowledgeable than their younger contemporaries; however, students who have overstayed their expiration date are likely weaker students than those who have finished in a time period commensurate with expectation. The two-way interaction suggested here is in keeping with the results presented in 4.2.2.3, but is beyond the scope of this dissertation and thus will not be investigated further at this time.

4.3 Summary of Research Question 2

The second research question, *What do the results of the KPVMI-1 administration tell us about the current level of p-value fluency among doctoral students nationally?,* was addressed (considering the sample limitations) with an analysis conducted on the Vignette subsection of items only (*KPVMI(P)).* This decision ensured that any implications drawn about the p-value fluency of doctoral students nationally was based on the most reliable subset of available items. In general, the median score was 10/18 items correct, or slightly better than half (56%). Small, but wholly insignificant, differences in mean performance were seen in the subgroup comparisons for Sex/Gender and Major. By subscale, ASA principles #1 and #6 appear to be the most elusive and #3 and #4 appear to be the most easily understood – this trend is essentially consistent across respondent characterizations by Major, Experience, and Status classifications.

Chapter 5 – Discussion

In this chapter, a discussion of the results will be presented, including answers to the research questions. In this discussion, a final assessment of the current study will be offered which addresses the quality of the effort and the fidelity of the results. Provisional conclusions, within the boundaries afforded by the study limitations, are reported with an eye to potential contributions to the research literature base. The chapter closes with the proposed future directions for this research.

5.1 Discussion of Research Question 1

The first research question, *Can a sufficiently reliable and valid measure of p-value misinterpretations (in a research context) be constructed*?, was addressed via the development and validation of the *Keller P-value Misinterpretation Inventory* instrument (*KPVMI-1*). The pursuit of the first research question can be thought of as involving three primary objectives: item development, instrument construction, and instrument validation. As a point of clarification, the three phases of the study (Phase I – items design, internal & external models, construct map; Phase II – pilot test; Phase III – field test) do not directly correspond with these objectives in a linear progression but rather necessarily overlap, traverse, and revisit them as the iterative process dictated. In this discussion, each of the three objectives will be discussed in sequence, drawing on evidence from any and all applicable phases of the study when appropriate to the conversation. Cautiously, it can be reported that there was evidence of satisfactory progress towards all three objectives; however not necessarily *complete* satisfaction, as this discussion will elucidate.

5.1.1 Discussion of Item Development

In any instrument development endeavor, it is important to remember that the “instrument is always something secondary”; as Wilson explains, “there is always a purpose for which an instrument is needed and a context in which it is going to be used" (p. 6). Thus, this branch of the discussion naturally begins with an evaluation of the item development process, paying specific attention to how the foundation for the instrument was laid with the theoretical framework (i.e., the construct map and internal/external models).

5.1.1.1 Theoretical Framework

Upon reflection, the situation of *p*-value fluency in the external model remains an appropriate decision. Declaring *p*-value fluency a necessary component of PhD scholarship dictated the context for the instrument (i.e., PhD students) and the purpose (as a means of assessing the potential for future researchers to transition into contributing members of the research community). The uncertainty lies within the other building block components.

By design, the composition of the internal model should be a direct consequence of the construct specification and external model. In that regard, this research has acted accordingly; however, the initial choice of, and subsequent modifications to, the construct domain must be defended. An instrument, again according to Wilson, is the "result of a series of decisions that the measurer has made regarding how to represent the construct or, equivalently, how to stratify the 'space' of items" (Wilson, 2005, p.45). In this study, Greenland et al.'s list of *p*-value misinterpretations was selected as the basis for the 'space' (or content domain) on the grounds that their list was based on identified issues *in press* and not based solely on research performed on students (specifically undergraduate students or students in a first course in statistics, in most cases). Furthermore, its comprehensiveness seemed to subsume any and all viable alternatives suggested in the research literature. Upon reflection, new evidence has come to light that calls into question the temerity of this position.

Greenland et al's list consisted of 25 misinterpretations subdivided into four classifications: misinterpretations of single p-values, misinterpretations of p-value comparisons and predictions, misinterpretations of confidence intervals, and misinterpretations of power. The initial proposed internal model reflected this structure, as well as the specifications of the external model, by diagramming *p*-value fluency as consisting of four components each in two process levels (independent research: out of context, peer review: in context). The inherent assumption underlying this presentation is that the construct (*p*-value fluency) *necessarily* contains *all* of these concepts. It is taken as a limitation of this study that the instrument ultimately failed to reflect this structure. Uncertainty regarding the true dimensionality of the construct, coupled with concerns regarding proposed response rates, resulted in the decision to eliminate the latter two categories of misinterpretations from the construct domain when populating the item pool (resulting in 18 misinterpretations distributed across 2 categories).

This decision turned out to be prudent[48] for a variety of reasons, discussed here in no particular order. Firstly, the item development and validation (with regard to lexicon specifically) proved increasingly difficult as the concepts increased in technicality and obscurity. The latter two categories, confidence intervals and power, were likely to have presented a challenge in this regard that might have been unable to be resolved in the phases of data collection planned. Secondly, these latter categories appear (although there is no empirical evidence in this study to validate this assertion) to contain advanced ideas that a respondent would have trouble negotiating without a strong fundamental grasp of the basic definition and interpretation of a *p*-value. In other words, these topics *seem* to be situated on the upper limit of the continuum as they involve the coordination of multiple related constructs simultaneously

[48] *Prudent* in the short term, but it is acknowledged that it might not have been *correct* in the long run. This is discussed further later in this chapter.

(e.g. an understanding of the *p*-value in the context of an item about power combines ideas of the hypothesized distribution, the sample size's effect on that distribution, effect size, and possibly error rates) and are thus likely out of reach of the majority of respondents. Field test results seem to support that hypothesis as the median performance level was a mere 10 of 18 items answered correctly[49]. Respondents who struggle to answer the *easy* items, would undoubtedly struggle with these more advanced items – the consequences of which would lead to lower scores, increased item difficulty indices (affecting validation), and possibly increased attrition. Finally, the aforementioned potential issues surrounding dimensionality were in fact realized. The inability to definitively establish the process levels and number of underlying dimensions of the construct would surely have been muddled even further with the inclusion of additional categories and overly difficult items.

5.1.1.2 Item Technical Quality

Accepting our choice for the theoretical basis (dismissing the potential inappropriateness of that decision for now), the discussion will now address the technical quality of the items themselves. The original wording of the context free (T/F) items matched verbatim the wording of Greenland and his colleagues. Based on input from the item panel, these (false) statements were abbreviated and/or modified to reflect language familiar to the target population and to present stems as free from ambiguity as possible. When drafting the inverse (true versions) statements, consultation with the item panel was employed in order to isolate the exact underlying idea that was to be nullified or negated and then all reasonable efforts were taken to construct the stem as analogous to the original (false) version as possible. The expertise of the panelists provided valuable insight into both the concepts underlying the misinterpretations as

[49] The author acknowledges that this level of performance could be attributed to *either* the difficulty of the items or the skill level of those sampled, but identifying which is indeterminable at the present time.

well as common student heuristics and conflations associated with such ideas. The pilot study (think-aloud interviews) provided persuasive evidence that the items were primarily functioning as desired and that sufficient technical quality had been achieved. Notable exceptions were flagged for revision and updated between the pilot study and field test rounds of data collection. No additional empirical evidence was collected to corroborate the technical quality of these revised items, but this was investigated during the instrument validation process (e.g. item discrimination indices).

Another valuable contribution of the item panel and the pilot study was in identification of the appropriate research scenarios to present in the vignette section. *P*-values have universal interpretations that do not rely upon the nature of the testing situation; that is, the practical implications of a hypothesis test may change when comparing (for instance) a 1-way ANOVA test of independent means with a *z*-test of a single proportion, but the true meaning of the *p*-value remains consistent across these scenarios. To illustrate with one of the identified misinterpretations, the *p*-value is *never* the probability that the null hypothesis is true regardless of which particular null hypothesis has been specified.

That being said, there was the issue of whether respondents would be equally familiar with interpreting *p*-values under different testing conditions and whether this would interfere with their ability to convey their level of understanding. On the one hand, the case could be made that failure to recognize the stability of interpretation across testing scenarios is in and of itself indicative of a lack of *p*-value fluency; however, that point had to be weighed against the potential influence that unfamiliar tests (i.e., ANOVA, Chi-square vs. the more ubiquitous *t*-test) might have on the test taker. After all, the purpose of this assessment was not to determine the respondents' breadth of familiarity with statistical tests, but rather to determine their *p*-value fluency within the context of a given test. If the interpretation of the *p*-value is truly universal and immutable across testing scenarios, why not provide the opportunity for participants to test under conditions in which they are more familiar?

Ultimately, ANOVA and Chi-square tests were removed from consideration in favor of the traditional one-sample and two-sample *t*-test of population means. This strategy was raised by the item panel and substantiated in the cognitive interviews. Participants verbalized concern when faced with unfamiliar scenarios and appeared to devote substantial mental energy to remembering what they had learned about these tests and to adjusting their frame of reference to 'think' in these alternate scenarios – as if preparing to answer questions about *ANOVA* testing as opposed to questions about *p-values* in an ANOVA context. This seemingly misleading and unnecessary increase in cognitive demand was most easily eliminated with the elimination of these items.

An interesting observation of this research was this notion that *p*-value understanding was somehow enmeshed with testing context. A cursory examination of the participants' free-response *p*-value definitions seemed to support the idea of this enmeshment as many responses made reference to a specific test or distribution in their definition. There is insufficient evidence to investigate this further in this research, but it is possible that group differences by Major were related to the choice of testing scenarios specified. Different contexts lend themselves to different tests and perhaps students in certain degree programs encounter particular tests more regularly than their counterparts in different programs.

5.1.1.3 Item Format

The final issue related to item technical quality that will be discussed here is in the designation of item formats. The use of a T/F format was controversial for a variety of reasons, many of which have already been explained in this paper. Apart from the general disdain for this particular format expressed by the item panel, there were the more specific issues of suitability to this content and contribution to instrument validity. Given that the basis for the content domain was a list of *mis*interpretations, that established an initial item pool consisting entirely of false statements. This was unsuitable to a T/F format, in which respondents would expect the correct answers to be somewhat evenly distributed amongst both categories, and necessitated augmenting the item pool to include true versions of each statement. Difficulties and details of this process have already been discussed at length in Chapter 4 and will not be repeated here, but suffice it to say, that striking the correct balance of T/F with the correct version of the negation so as not to misrepresent the underlying sentiment of the original misinterpretation was a complicated process.

Beyond the complications that accompanied the need to negate false statements into truths, there was the issue of how the T/F subsection affected the overall instrument validity. As discussed in Chapter 4, the pilot study cast speculation on the T/F section's ability to function appropriately. The primary concern was that in a traditional T/F format, the underlying expectation (as voiced by pilot study participants) was that the items were assumed to be true until evidence was found that negated this position. Thus, participants were searching for key words and phrases that *made the statement false*. This was problematic because it primed some of the participants to be suspicious of *trick* questions. Furthermore, it caused the participants to impose undue influence on particular words and their implied meaning (e.g. does *upheld* mean the same thing as *prove*?, does *confirm* mean the same thing as *corroborate*?, how to interpret *absolute truth*).

Tangentially related to this inability of participants to definitively assign truth (or falsity) to particular statements was the shades of grey that infiltrated what should have been a purely black and white exercise. While some of the original statements were patently false (and statistically or mathematically verifiable), some appeared to be ruled by convention in lieu of statistical theory. The p-value, by definition, does *not* tell you the probability that the null hypothesis is true (misinterpretation #1). There is no room for negotiation here. Contrast that with misinterpretation #8 (*Lack of statistical significance indicates that the effect size is small*.) in which *small* is open to interpretation or misinterpretation #12 (*p-values are properly reported as inequalities*.) in which one's belief in this assertion is as much a function of his/her discipline as it is a reflection of *p*-value fluency or statistical training. Different journals use different reporting standards and respondents accustomed to seeing the star method ($p < .05$ = *, $p < .01$ = **, $p < .001$ = ***) in print might conflate *accepted* practice with *best* practice.

The introduction of the P/CP format was meant to alleviate these concerns. In this presentation, respondents no longer needed to concern themselves with the myriad explanations as to why a particular statement might be false, but rather select the *better* reporting option from only two alternatives in which the primary distinction between them had been unmistakably denoted. In the case of black/white items, the test-taker knows that one of the statements must be true and that the *correct* answer is listed. In the case of *grey* items, the P/CP format permits the items to differentiate between *shoulds* and *coulds* and allows the respondent to dually consider those scenarios with which they are familiar (convention, accepted practice) against a purportedly superior alternative (expert recommendation, best practice) without necessarily undermining or maligning one's support or belief in the former.

The improvement to instrument quality and reliability that occurred upon replacing the T/F section with the P/CP section supports the decision to do so. That the T/F items were merely inserted verbatim into the P/CP section (in most cases) as opposed to being recrafted for the new format corroborates the idea that the problem was not with the technical quality of the item stems, but with the item format – both in this particular instance as implemented and in general when applying this item type to this kind of content (i.e. *mis*conceptions). While the P/CP subsection failed to stand alone as a complete 18-item subscale upon validation, the success of the vignette section (also in P/CP format) suggests that this format is viable and that the appropriate course of action is to tweak the individual items exhibiting poor discriminations or model misfit (as opposed to abandoning both the T/F and P/CP format entirely for context-free items).

5.1.2 Discussion of Instrument Construction

Constructing a test blueprint is a function of many factors. In this section, the following components will be discussed: test coverage, test length, and item formats.

5.1.2.1 Test Coverage

The test blueprint is ideally a reflection of the internal model of the construct. As Wilson described, the instrument is the product of a decision to stratify the “space” of items and sample from within those strata. This process is iterative and inevitably circuitous – when done properly. In the beginning, “the construct and items are both only vaguely known…[and] causality is often unclear” (2005, p. 11-12). When a specified relationship is presumed and maintained without an evidentiary basis, “this unfortunate abbreviation of the instrument development typically results in several shortcomings: (a) arbitrariness in choice of items and item formats, (b) no clear way to relate empirical results to instrument improvement, and (c) an inability to use empirical findings to improve the idea of the construct" (p. 12).

Inspiration for an instrument can come from either direction. Sometimes the researcher has a vague notion of a construct and then attempts to write items that capture it; in other cases, "the items will be thought of first and the construct will be elucidated only later" (Wilson, 2005, p. 10). In this research, a rather convincing argument could be made for either situation, and in reality, they were both true. Controversy surrounding the application and interpretation of *p*-values in practice certainly conjured to mind the idea of *p*-value fluency as a construct worth defining and measuring. Conversely, the list compiled by Greenland, et al. (as well as the other lists cross-tabulated in Table 2) presented as an initial item pool from which a construct could be built. That the misinterpretations were already conveniently arranged in sub-categories was seductively deceptive for its potential to masquerade as a fully defined construct, but to never investigate beyond this assumption would have been exactly the inappropriate abbreviation to which Wilson was referring.

In the previous section, the discussion raised the issue of the validity of the internal model, not in a way to draw attention to any inappropriate researcher action but rather to explicate and emphasize the unfinished status of this endeavor. An initial content domain was selected and the internal model, construct map, and test blueprint were generated to match. As empirical evidence from the item panel and pilot study (and eventually, field test) informed the construct, these artifacts were necessarily updated. At present, the understanding of the construct is that it contains at least 18 misinterpretations that are measurable at up to 2 process levels. Whether both process levels (in/out of context) are necessary to fully define the construct is yet unknown. Whether the list of 18 misinterpretations is sufficient (or superfluous) to capture the construct fully is also unknown.

Every instrument design process must designate a starting point, an initial content domain that may or may not match the ultimate test blueprint and construct map. In this study, neither the choice to use Greenland et al.'s list nor the decision to reduce it were whimsical as prudent justification (already presented in this section) existed; however, these decisions were somewhat arbitrary. Arbitrary in the sense that if one is committed to investigating any and all possible subsets of items and contexts to infer the underlying construct, the choice of starting point is somewhat arbitrary if one intends to traverse the entire map eventually anyway.

When assessing the quality of this research, it is both appropriate to acknowledge that the current presentation of the internal model, construct map, and test blueprint are properly aligned and valid for the *present* understanding of the construct and to concede that all are likely incorrect because they are based on an incomplete and inaccurate model of the construct. In other words, the present instrument construction is an appropriate reflection of the construct as presently understood, but will require modification as the data informs the model of the construct moving forward.

When presenting these artifacts (i.e., internal model, test blueprint), the inherent underlying assumption is that the construct necessarily contains *all* and *only* these concepts; however, this declaration is only appropriate when final. It is a limitation of this study that it was neither able to provide data on respondents' capabilities in the supplemental categories (misinterpretations 19-25) nor able to provide compelling model-based evidence justifying their exclusion – but only a limitation in the sense that the cycle was prematurely arrested without having sufficiently mapped the construct. In other words, the internal model was updated to match the test blueprint, but without a strong sense that either was correct.

5.1.2.2 Test Length

All things being equal, reliability favors a longer instrument. From a measurement standpoint, having replicate measures within subscales and contexts is desirable for achieving more stable and valid parameter estimates; however, that agenda must be accomplished in such a way as to keep the total test length abbreviated enough to assuage concerns of redundancy and test-taker fatigue.

The pilot version of the instrument was the longest by design. At this early stage of data collection, the item pool was still quite large and a variety of item formats and contexts were still being investigated. A primary purpose of this round of data collection was to weed out poor items and thus was purposefully extensive. Additionally, as these participants were being paid for their participation, their cooperation, even under a longer test condition, was a reasonable expectation.

In each subsequent round of testing, the instrument was abbreviated based on an array of factors. The most important determining factors were item performance and user feedback. When the T/F section was eliminated due to performance concerns, this shortened the test from 3 sections to two and reduced the number of items by five. When the even vignette section demonstrated inferior performance to the odds, this section was eliminated reducing the number of items by 18, but also significantly reducing the reading burden with the elimination of the four affiliated vignettes. This decision was made not solely on the basis of performance, but also due to user feedback – consistent among the comments was one or both of the following sentiments: the test is too long or the items are too redundant (see Appendices T & Q for details).

While some of the redundancy was by design, i.e., testing the same misinterpretation in both contextual and context-free settings, there was perhaps no good reason to test the same misinterpretation repeatedly in either process level (other than possibly from a measurement standpoint), especially considering the incentivization available. Increased test length would lead to lower response rates and lower completion rates – both of which would reduce the sample size. Again, a balance had to be struck between a large number of responses consisting of a few items or a small number of responses consisting of many items.

A principle consideration of any test length decision is the intended setting for administration and the potential consequences to participants; the higher the stakes, the greater the need for precision and accuracy. The purpose of this instrument is to measure graduate students' p-value fluency as a statistical disposition is pertinent to their future careers. Under the assumption that the assessment could discriminate between those students who are capable of applying the ASA principles appropriately and those who need more preparation in this aspect of their plan of study, the intended application is that this instrument could be used in conjunction with one's qualifying/preliminary exams or as an exit assessment for quantitative methods coursework. In this context, the stakes are high and the participants are invested – test length would likely be wholly irrelevant. The trouble is that the development of the instrument takes place in a different context where the participants are not necessarily stakeholders in the outcome and are commonly not properly incentivized, thus their tolerance for fatigue and redundancy is comparatively diminished. Perhaps the most sensible solution is to generate parallel, reduced-length, forms of the instrument whereby different participants are testing different subsets of the total item pool. This way, data is being gathered for all items, but the burden of items/participant is manageable. A similar strategy was employed in Phase IIB of this research as the cognitive lab participants were assessed on either the odd or even vignette items, but not both.

The current understanding of the construct and validation efforts have suggested that an 18-item instrument in which each misinterpretation is tested exactly once in a contextual setting might be appropriate as a stand-alone measure. Nevertheless, future research in determining the necessity of multiple process levels (in/out of context) and the sub-scale structure (i.e., by category, by ASA principle, by otherwise unidentified factor structure) is necessary before declaring a final test length.

5.1.2.3 Item Format

A thorough discussion of the viability of the T/F and P/CP formats has already been presented in Section 4.1 and will not be repeated here, but there is another important formatting issue that still warrants further consideration, namely the addition of a third response choice for all items. In the current format, the respondents are forced to make a selection for each item. All conclusions drawn regarding the accuracy of such selections assumes that the participant intended to make that selection on the basis of an inherent understanding of the underlying concept or its realization in the particular item. This is a limitation of using this type of item.

In the cases where a person has a strongly held conviction in either direction, the binary format is ideal; however, without interview data, there is no way to filter out the guesses from the meaningful selections. In some cases, a participant does not have enough knowledge of the subject to understand the item or differentiate between the alternatives; a third option reveals that inability to properly engage with the content. In traditional knowledge assessments, it is perhaps irrelevant whether a participant answers incorrectly because of a conviction in an incorrect solution method or due to ignorance, the end result is an incorrect answer. In the context of this study however, it is important to distinguish between misinterpretations and misconceptions (e.g. *I think the p-value tells me the probability the null is true)* and ignorance (*I don't know what an effect size is).*

After the pilot study, and before the field test, it was proposed to institute this modification to the item format. Transcripts from the think-aloud interviews revealed that participants responded to unknown items in a variety of ways, sometimes illogical or idiosyncratic. One respondent would stipulate that a particular choice "made sense to her" and then proceed to select the alternative, as if to imply that she thought the more complicated option must be the correct one. In another case, a respondent would make a selection based on stem length. There is also the issue of respondent tendencies. Some participants would always select *Point* (for instance) in the absence of a genuine selection. These selections are in contrast to when participants would verbalize conviction in a particular misinterpretation and (appropriately) select the incorrect answer. Without a third option, the response pattern is as much a function of luck or instrument construction (e.g. more 'Point' answers are correct) than it is of participant capability.

In the context of this instrument, and the intended applications, there was arguably justification for filtering the *p*-value concepts that people *mis*understand with those that are unfamiliar. The former situation requires remediation whereas the latter requires instruction. This was debated by the item panel and ultimately this change was not instated. The primary concern was that the third option would be a "dumping ground" for respondents and would lead to sparse data that would not contribute to item parameter estimates.

Interestingly, many participants decried the lack of a third option. This was a frequent complaint/suggestion aired in the comments section of the exit interview. Ultimately, there was nothing methodologically inappropriate in sticking with the binary format; however, this research will keep an open mind about investigating the usefulness of such a strategy in future iterations of data collection.

5.1.2.4 Instrument Usability

Throughout all phases of this research, usability data was collected in order to construct an instrument that was user-friendly. Details such as font size, number of items per page, arrangement of distractors, presence of progress bar, ease of accessibility to survey link, and even the ordering of the test sections were addressed in accordance with participant feedback. The usability of this instrument has been deemed satisfactory. Participants are not generally dissuaded by the projected completion time, have no difficulty accessing or navigating the survey, and largely tend to agree with the formatting.

Upon review of the most recent round of exit interview data (field test, Appendices Q & T), participants raised two issues that warrant further consideration. The first of these was the absence of a *Back* button that would allow respondents to navigate freely between sub-sections of the instrument. This option was contemplated briefly but abandoned due to the redundancy of the instrument. Later items tended to reveal answers to earlier items (specifically vignette scenarios sometimes influenced and caused modifications to the context-free items) and the item panel felt it best to avoid this scenario if possible. Future versions of the instrument, depending on the number and type of items, might not suffer from this affliction. If, and when, such time occurs, the option of free navigation can (and will) be reexamined.

The other of the participant suggestions was a scoring mechanism. Respondents indicated a desire to gauge their performance, both in total and by item. While there is danger of contaminating future data by publishing the answers, there does not appear to be a sufficiently good reason not to inform the respondents as to their performance, at least overall. Individual item feedback would be akin to releasing an answer key, but sharing total scores and subscale scores (if appropriate) seem to offer more benefit than harm. The fact that people are interested in this information suggests that having their consciousness raised about the misuse of *p*-values has caused concern as to one's group membership (i.e., *Am I one of those who are doing it incorrectly?*). Furthermore, receiving a low score might actually motivate that individual to seek clarification and/or additional instruction. In the absence of scores, there were already some individuals that indicated an intention to do precisely that in reaction to their perceived performance. Any contribution to the *p*-value conversation, no matter how slight, should be taken as a positive consequence of this research – if releasing scores will achieve that, then perhaps it is a modification that should be given legitimate consideration in future rounds of data collection.

5.1.3 Instrument Validation

The instrument validation plan for this project was devised in the Messick tradition, structuring the validity argument around six dimensions: Content, Substantive, Structural, Generalizability, External, and Consequential. A thorough discussion of this plan as enacted was presented in Chapter 4 and will not be repeated here. A summary will be offered but merely as a means to facilitating the present discussion which will address the quality and fidelity of the validation effort.

On the whole, the argument could be made that a reasonable validation plan was proposed and satisfactorily implemented. Validity evidence was able to be gathered for at least one measure in all six dimensions; although, it must be acknowledged that there was no dimension for which data was collected for *all* indicators and occasionally, a proxy was necessary when a suitable direct measure was unavailable. In that way, there is room for improvement as the instrument development progresses.

According to Wilson, "the gathering of evidence should not be seen as simply a once-and-for-all event" (2005, p. 156). In that regard, this research wholeheartedly embraced that notion. Validity evidence was collected at every stage of data collection and used to inform the next iteration. Even when data collection had ceased, an action plan to improve the instrument beyond the scope of this dissertation was proposed and implemented to the extent that it was possible.

The validation efforts were *successful* if measured by the more commonly reported measures such as *Coefficient Alpha* and *Item Discrimination*. The PV3 version of the instrument (30 items assessing 18 misinterpretations across 2 process levels) exhibited a reasonable *alpha* = .8030 with acceptable item difficulty and discrimination indices for all items. The *KPVMI(P)* version of the instrument, functioning as an independently validated sub-measure of the construct (18 items assessing 18 misinterpretations at a single process level) exhibited a slightly improved *alpha* = .8298 with, again, acceptable item difficulty and discrimination indices for all items.

Within each of the six dimensions, there was positive evidence – such as the indicators just named – that the instrument was functioning as desired (see also Table 41); however, it would be shortsighted to discontinue validation efforts on the basis of this *success*. As Wilson describes, sometimes the negative evidence can provide valuable information: "The purpose of validity evidence in instrument development is to help the measurer make the instrument work in a way that is more consistent with the intent, and evidence that the instrument is not doing so is not a dead end in this process" (p. 156). Negative evidence identifies problems, the resolution of which improves the quality of the instrument.

In this research, negative evidence was accumulated in five of the six dimensions (Consequential being the lone exception). There are "at least three different sorts of conclusions one can draw from negative evidence: (a) the original idea of the construct (theory of the construct) was in some way wrong…(b) the items that were developed to match the construct are not working as intended; or, (c) the scores that have been developed for the items are incorrect…"(Wilson, 2005, p. 161). The third option is not really applicable here as person-fit data was inestimable, but the other two scenarios can be discussed.

The implications of negative evidence, and the corresponding revisions therein, are already well-documented in this paper, but will be listed here in brief for illustrative purposes. Negative feedback from the item panel regarding the technical quality of the items and formats resulted in revisions to both the item wording (b) and the construct map (a). Negative feedback from the think-aloud interviews regarding the readability of the vignettes, the clarity of the instructions, the vocabulary of the items, and the formatting of the sections resulted in revisions to all (b). Additionally, negative feedback regarding the familiarity/comfortability with alternate testing scenarios informed the content of the vignettes (a). Negative item performance of the T/F section resulted in the elimination of these items (b). Negative feedback from the IRT modeling and factor analysis resulted in proposed changes to the construct map and internal model (a).

Arguably, there is sufficient substantiation for the claim that both positive and negative validity evidence were properly gathered and addressed. Where this research is deficient, is in those dimensions for which evidence was neither positive nor negative but instead missing. In the External and Consequential validity dimensions, there were indicators that were inestimable.

Specifically, the absence of person-fit data and the absence of a true convergent validity measure are conspicuous omissions that must be taken as a limitation of this research and addressed as this study moves forward.

In the Content, Substantive, and Structural dimensions, the evidence was not missing completely but incomplete. A largely unresolved issue of this study is the uncertainty surrounding the definition of the construct (as explained in detail in 5.1.1.1 and 5.1.2.1); in other words, the failure to crystallize the realization of the construct. As Section 5.3 will explain, the limitations of this research can very nearly all be traced to an insufficient sample, but it is important to revisit these missing validity components when new data is in hand. While all dimensions of validity matter, "evidence based on instrument content is *essential* because it contains the realization of the construct, and that is what all the other aspects of validity play off" (Wilson, 2005, p. 157, emphasis added).

5.2 Discussion of Research Question 2

The underlying motivation for this research was to investigate whether the controversy surrounding *p*-values and their misuse in practice had generated enough attention for the community to chart a course of self-correction. In other words, would the next generation of researchers (i.e., current doctoral students) surpass their predecessors (i.e., current researchers) in that regard; hence the second research question (*What do the results of the KPVMI-1 administration tell us about the current level of p-value fluency among doctoral students nationally?).* The explanation for that being the *second* research question is the non-existence of a sufficient instrument for conducting such an investigation. In this section of the paper, as instrument validation remains a work in progress, we endeavor to answer that question in the limited capacity that the size and nature of the sample permit and using an abbreviated version of the instrument, the aforementioned *KPVMI(P)*.

The second research question was addressed via analysis of a subset of the field test data (n = 147) with respect to performance on the subset of items considered sufficiently validated as developed in Phases I-III (*KPVMI(P)*). Appropriate descriptive statistics for overall performance, and by sub-category of misinterpretations where factor analysis indicated such a calculation was warranted, were performed. Secondary analysis compared performance figures by subgroup (e.g. by gender, by program of study, by level of preparation, etc.) and made inferences therein. Specifically, this phase of the research investigated the following sub research questions: *Are there performance differences by subgroup? Can these differences be attributed to differences in experience or preparation? What are the implications of these results in the context of the ASA's principles?*

The discussion of the results will be framed around identification of potential contributions to the research literature while being careful to differentiate between what *can* and *cannot* be inferred from the findings. Note that any generalizations implied or expressly stated in this discussion are based on the current sample and are not necessarily nationally representative. Appropriateness of extrapolation of the results beyond the current sample is discussed in detail in Section 5.3 of this document.

The purpose of research question two was to collect baseline data on the level of *p*-value fluency in doctoral students nationally. Assuming for the moment that the *KPVMI(P)* is in fact a legitimate measure of that construct, the results indicate that the concerns raised in the research literature regarding this ability are well-founded.

5.2.1 Performance Results Overall

According to Greenland and his co-authors, each item on their list has “contributed to statistical distortion of the scientific literature” (2016, p. 340). This list was compiled based on existing research (i.e., reflects the understanding of practicing researchers) and an important objective of this dissertation was to investigate whether future researchers (i.e., doctoral students) suffer from the same misconceptions. It was hypothesized that perhaps the renewed attention to the *p*-value conversation would have contributed to an emphasis in methodological training in general, and *p*-value usage in particular, in degree programs for PhD students. There was a suspicion, if not an expectation, that perhaps the *future* researchers might be better equipped where *p*-value fluency was concerned than the previous generation (i.e., *current* researchers). The results do not support this hypothesis.

The average performance was 10/18 items correct (mean = 52%, trimmed mean = 57%, median = 55%). There was a sizeable amount of variance in item difficulties (i.e., proportion of correct responses), but no single item was *so easy* that the argument could be made that this list of 18 should be reduced. 82% of respondents answered the easiest item correctly compared with 38% of respondents answering the hardest item correctly. While the argument can be made that #12 is much more readily understood than #15, there are still nearly 20% of respondents who suffer that easiest misconception. In other words, there is evidence that future researchers as a group suffer from *all* the same misinterpretations as practicing researchers do, albeit to varying degrees.

5.2.2 Performance Results by Subscale

When looking by subcategory, the *Common misinterpretations of single p-values* items unsurprisingly seem to be more readily understood than the *Common misinterpretations of p-*

value comparisons and predictions items as judged by their average item difficulty indices (.5355, .4745 respectively). This discrepancy is perhaps not as large as one might expect but can be at least partially explained by the inconsistency in performance among items that would be expected to cluster together based on conceptual similarity/proximity (as discussed in Section 4.2.1).

This binary categorization offered by Greenland is but one of many possibilities for sub-classification and not the only one explored in this study. As an alternative, this research proposed a correspondence between the misinterpretations and the ASA principles for *p*-value usage. When looking at the items grouped in accordance with the ASA principles, the respondents seemed least likely to be susceptible to violations of Principle #3 and, in descending order, Principle #4, Principle #5, Principle #2, Principle #6, and finally, Principle #1[50]. This alternate classification is still based on the same set of 18 misinterpretations and thus does not suggest a different result than the previous one: future researchers as a group struggle in *all* of these areas.

The results might not suggest an entirely different conclusion when categorizing the items by ASA principle; however, what is a useful contribution of this classification scheme is in how the results can inform the education and remediation of these concepts. What Greenland's categorization fails to provide is a clear roadmap to amelioration of misinterpretations. Broadly separating misinterpretations as being used in either a *single* or *comparison* context doesn't really pinpoint the etiology of the misuse/misunderstanding. To be fair, the *comparison* group is fairly homogenous and rather small (only 4 items), but the *single* group is not and thus could

[50] When discussing these results, it is important to bear in mind that the subscales are not comprised of equal numbers of items. The reliability of these measures, and thus any implications drawn therein, are not necessarily robust to this inequality and likely to stabilize as the number of items increases (for instance, the subscore for Principle #3 is likely the least trust-worthy as it is the only subscale consisting of a single item).

stand to be improved with further sub-classification. Within the 14 *single* misinterpretations, there are at least 3 clearly distinct subcategories that can be identified: statements for what the *p*-value *is* (definition-related), what the *p*-value *implies*, and how the *p*-value should be *reported*. Factoring in peripheral ideas like effect size and Type I errors, this could be extended even further. The ASA principles, as a guide to "[improving] the conduct or interpretation of quantitative science" (Wasserstein & Lazar, 2016, p. 131), have composed a list that reflects the intricate nature of *p*-values in recognizing that proper use necessarily entails all of the following: understanding of the definition of a *p*-value, recognizing the limits of the implications and inferences *p*-values provide, and adhering to the expectations for reporting a *p*-value.

Collectively, the group of respondents suffers from misinterpretations in all six ASA principles, but individually, it is quite possible that personal profiles show success in some particular aspects of *p*-value fluency. As a diagnostic tool, it would be valuable to be able to identify the specific principles that are in need of attention. This is true on an individual level in the context of using an instrument like this in conjunction with qualifying/preliminary exams or a methodological coursework exit assessment, but perhaps it is of even greater value on a group level. Goodman, in reference to the ASA *p*-value statement, remarked that "we need to formulate a vision of what *success* looks like…" (Matthews, 2017, p. 40), and that is what distinguishes the ASA principles. Unlike the majority of *p*-value research in which misunderstandings, misconceptions, and misinterpretations are the focus, the ASA statement is written in positive terms – in other words, it doesn't describe all the ways people get it wrong, but instead depicts what doing it right would look like. In assessing the efficacy of statistical preparation at the program, department, institution, or even discipline level, it would be valuable to know which of the principles are/aren't showing satisfactory performance because changes could then be imposed at the curricular level.

The underlying dimensionality of the latent construct, *p*-value fluency, has not been established with this research and it therefore might be tempting to conclude that under a unidimensional assumption, it would not be possible to separate performance along subscales (ASA principles, Greenland's categories, or any other organization). The results of this sample indicate otherwise. In the event that the principles (or other designated classification) were highly related, enmeshed concepts in which competency in a particular area was difficult to extricate from the others, we would expect to see evidence of multicollinearity and/or strong bivariate correlations.

Examination of the correlation matrix (Table 5.1, cells display *r* with the *p*-value in parentheses) shows that the bivariate correlations between the six subscales (items clustered by ASA principle) range in magnitude from .084 to .392 – values consistent with weak to moderate relationships as defined by Cohen (1988). It is important not to be deceived by the statistical significance attached to some of these correlations and consider the alternative r^2 measure instead. Ranging from just under 1% to roughly 15%, this is the amount of shared variance between the pairs and keeps these relationships in the proper perspective. The strongest correlation is observed between #5 (*A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*) and #3 (*Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*); the weakest between #1 (*p-values can indicate how incompatible the data are with a specified statistical model.*) and #4 (*Proper inference requires full reporting and transparency.*). Based on the content of these principles, these orderings are consistent with expectation as the congruence of the former and unrelatedness/dissimilarity of the latter are fairly self-evident.

Table 5.1 – Correlation Matrix by ASA Subscale

Correlations					
	ASA1	ASA2	ASA3	ASA4	ASA5
ASA1					
ASA2	.218 (.034)				
ASA3	.310 (.002)	.293 (.004)			
ASA4	.084 (.418)	.243 (.017)	.254 (.013)		
ASA5	.266 (.009)	.264 (.009)	.392 (<.001)	.242 (.017)	
ASA6	.108 (.303)	.250 (.015)	.172 (.097)	.163 (.111)	.390 (<.001)

In some ways, this result is a bit counterintuitive. Even in the absence of a unidimensional structure, one would expect that the various components of *p*-value fluency would work in concert with one another, i.e., that the underlying concepts and ideas would not stand in isolation and would necessarily coalesce. It seems unlikely that a person could simultaneously hold both correct interpretations and misconceptions of the same central construct with no apparent cognitive dissonance, yet this appears to be the circumstance for the majority of respondents. As further evidence to advance this theory, there is reason to believe that not only is it possible to have inconsistent subscale performance (i.e., a fractured performance – high scores on some categories and low scores on others), but it is quite possible to exhibit correct understanding of interpretations and usage despite being unable to articulate a correct definition of the *p*-value itself.

5.2.3 Performance as Related to Understanding of the *p*-value Definition

As a supplement to the *KPVMI* items, a few additional questions were included on the instrument, one of which asked respondents to answer freely to the following prompt: *What is a p-value?* One of the intended uses for this data was as a proxy for a convergent validity measure assuming that people unable to define the *p*-value properly would likely score poorly on the assessment. The responses were individually scored by two raters (researcher and one of the statistical experts from the item panel) and consensus was reached by mutual agreement in the case of discrepancies. Scores ranged from 1 (“I don’t know” or similar), 2 (mostly/wholly incorrect), 3 (partially correct), to 4 (correct)[51]. Results can be found in Appendix AB.

Of 147 responses, only 14 were deemed acceptably accurate (these are highlighted in yellow in the table). This result in and of itself is somewhat immaterial and not all that surprising considering the proclivity of the research literature to report on this difficulty with defining the *p*-value (e.g., Goodman, 2008). What makes this result interesting, and ultimately germane to this discussion, is that knowledge of the *p*-value definition and performance on the *KPVMI* would seem to be independent. The distribution curves for total performance, as sorted by *Correct/Incorrect p*-value definitions (Figures 5.1, 5.2) shows that respondents in both categories flesh out nearly the entire spectrum of possible scores.

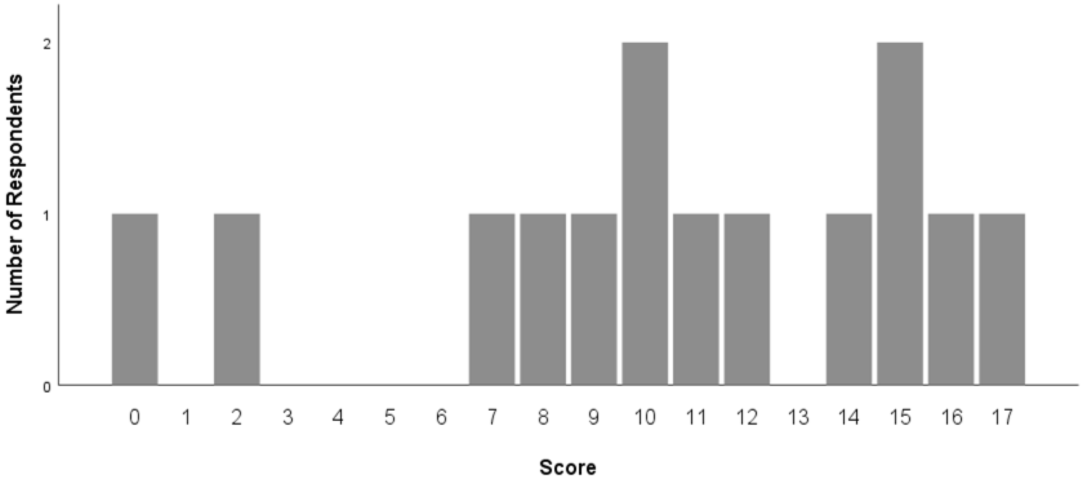


Figure 5.1. KPVMI-1 Score Distribution when a Correct *p*-value Definition was Supplied

[51] There was also a 0 score assigned to blank responses and a 5 score assigned in place of an asterisk. These responses were neither technically correct nor incorrect and referred to the p-value as merely “a probability” (which it technically is) or a measure of “statistical significance” (which it also technically is). As neither of these responses gave the formal definition, they could not be scored as correct, but what was said was not inaccurate and so could not be scored as incorrect; hence, the need for the * category.

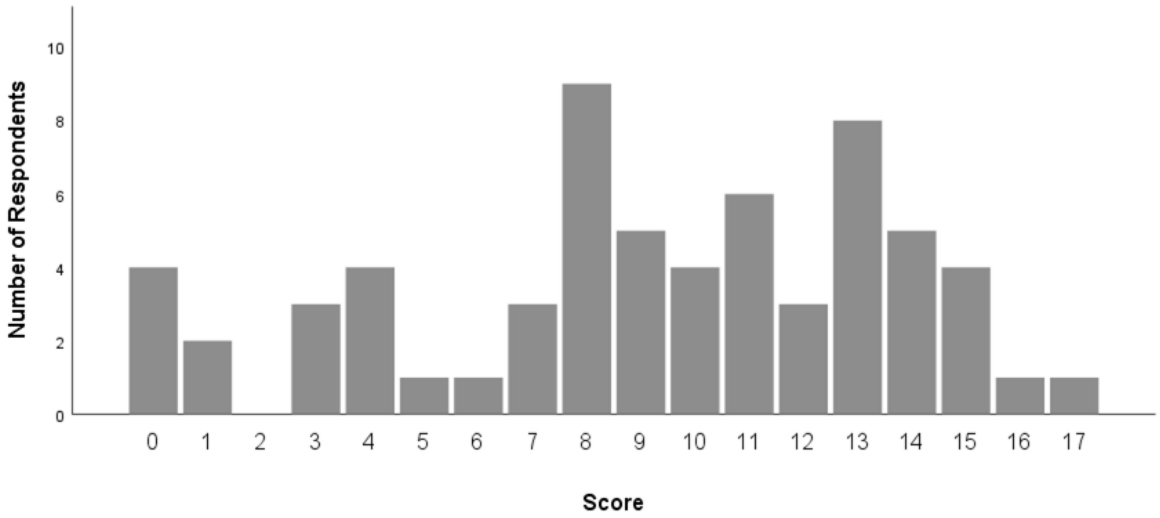


Figure 5.2. *KPVMI-1* Score Distribution when an Incorrect *p*-value Definition was Supplied

This independence is further corroborated with the Pearson correlation coefficient, *r* = .008. As a proxy for a convergent validity measure, these results would be *negative* evidence, i.e., an indication that the instrument was not functioning as intended. In this case, however, the totality of the validity evidence would suggest this anomalous result says more about the complexity of *p*-value fluency than about the invalidity of the *KPVMI* measure.

5.2.4 Constraints on the Implications of the Results

An important caveat when interpreting the results of the second research question is not to read into what the data might imply. The overarching objective of this research is to build an instrument to assess the level of *p*-value fluency in future researchers and potentially diagnose areas requiring remediation with an eye to improving the overall quality of the research community at large. It would be tempting to infer that satisfactory performance on this (or any equivalent) assessment would translate to methodological capability in the context of independent research or peer review; however, there is simply no precedent for such a claim with the present data. It seems logical that one could claim that respondents who score poorly (i.e., cannot recognize misuse of *p*-values in the hypothetical research vignettes) would similarly be unable to function competently in peer review (i.e., would not recognize the misuse of *p*-values in actual practice), but it is the reverse implication that is both desired and unattainable. Much as the *p*-value does not tell you the probability that the null is true – no matter how badly you want it to – the inference capability of the current data is restricted from making inferences of such scope. In the proposed context of qualifying/preliminary exams or methods coursework exit assessment, this instrument can merely identify those who are likely *not* qualified in this area, but does not imply that those who perform well are *sufficiently* qualified.

5.3 Limitations

According to the ASA's *Ethical Guidelines for Statistical Practice* (2018), "good statistical practice is fundamentally based on transparent assumptions, reproducible results, and valid interpretations" (p.1). When presenting the results of any research study, it is therefore important to reveal any and all impediments to overall quality and boundaries on the scope of the generalizations and implications that can be drawn from the results. The following discussion will present all known limitations and limiting factors [52]affecting this study.

Without hesitation, the quantity and nature of the respondents was easily the most limiting factor in this research. Not only were issues with sample size and quality a serious limiting factor in and of themselves, but as this discussion will demonstrate, they were the precipitating factor for the majority of the minor limitations.

The ASA declares an ethical statistician to be one who "employs selection or sampling methods…appropriate and valid for the specific questions to be addressed, so that results extend beyond the sample to a population relevant to the objectives…" (A2, p.2) and who "…includes appropriate disclaimers…when reporting analyses of volunteer data or other data that may not be

[52] This study makes a distinction between a *limitation*, something that affects generalization of the results, with a *limiting factor* , an impediment severe enough that it not only limits the generalization of the results, but also hindered or thwarted the planned methodological approach.

representative of a defined population" (B.9, p.3). In the spirit of such transparency in reporting, this paper attempted to employ appropriate selection and sampling methods across all stages of the study and disclosed when volunteer respondents were utilized. Adherence to the guidelines however does not automatically confer immunity against poor quality samples or low response rates.

5.3.1 Limitations of the Pilot Study

In the pilot test portion of this research, it was not actually the sample size that was a concern as the targets were either achieved (IIB) or exceeded (IIA), but rather the nature of the respondents. Apart from the obvious concerns that accompany a volunteer sample, there was the issue of diversity. All five think-aloud interview participants were recruited on the basis of degree completion progress and ability to provide elaborate and meaningful feedback, but they were also selected for their proximity and familiarity. These students presented as distinct (i.e., representing five different doctoral programs), but this was a bit misleading as all of them have received their methodological training with the same group of core courses taught by the same instructors at the same institution. In the end, the feedback was unique to the extent that could be expected from typical within-group variance; however, the notion that their common statistical upbringing did not influence their *p*-value understanding in a manner consistent with a *local effect* cannot be dismissed. Specifically to the extent that their feedback was used to make edits to the wording of individual items, the failure to include respondents from outside this localized group might have led to an inappropriate conflation of colloquial preference with universal standard (e.g. use of *null* hypothesis vs. use of *test* hypothesis).

This same concern did not extend to the usability study participants as they were recruited through the university-wide graduate student listserv. Any similarity in respondents would have been purely coincidental.

5.3.2 Limitations of the Field Test

In the field test portion of this research, sample size was problematic in both administrations (local and national). The local field test targeted students in a particular course (STAT 5616) selected on the basis of the presumed diversity and potential for cooperation therein. While the response rate was actually rather good by industry standards (79%? – check with Anne), the ultimate usable sample size was smaller than expectation. This turned out to be not just a limitation, but an actual limiting factor as this precluded the classical item analysis and IRT model parameter estimation portions of the instrument validation from being performed on this data. Furthermore, the nature of the population differed widely from expectation in terms of student status (i.e., year in school) as the proportion of Masters' students among the respondents was problematically high.

In the national field test, to say the sample size was problematic is an understatement – abysmal is perhaps more fitting. The target was 500 responses which seemed like a reasonable assumption on the basis that each of the 115 potential institutions need only average less than 5 participants each. Out of the hundreds, and possibly thousands, of doctoral students enrolled at these R1 institutions, five seemed legitimately attainable. In hindsight, there were many possible explanations as to why this target failed to be realized.

Although the sampling frame consisted of 115 institutions, the respondents at the majority of these schools never had the chance to participate. Due to the (surprising) and overwhelming success of the graduate listserv for recruitment in the local field test, this was the dissemination method selected for the national field test. Unfortunately, use of this method required access to the listserv which – for an outsider – was particularly difficult to obtain. Efforts to track down the executive board of the local GSA (or similar) were only moderately successful – partly because of the timing (before/during the first week of the semester) and partly because, in many cases, the most recent election results had not been updated online and the contacts were outdated. No response whatsoever occurred in over 58% of the cases.

Of those who did receive the materials, not all were willing or able to participate with local dissemination of the recruitment materials. This was an unanticipated consequence of using this delivery method. Failure to reach the target sample size was not as much a function of respondent unwillingness or disinterest, but rather the inability for the message to reach the potential respondents. The original idea was to contact departments and/or professors to ask for their help in recruiting potential participants, but this idea was rejected after observing how unsuccessful that tactic was in the pilot study. It was naïve to assume that what was efficacious (and not) at VT would be equally potent (or not) elsewhere.

Finally, there was the issue of incentivization. In the pilot study, participants were paid. In the local field test, participants received extra credit. In the national field test, paying every participant was simply cost prohibitive – especially on the order of potentially 500 individuals. Without direct access to instructors, and the timing being out of synch with most academic calendars, offering extra credit was both infeasible and immaterial. In the end, the only available option – raffle prizes – was simply insufficient inducement for the majority of potential participants. This was evident not only in the meager response but also in the completion rate. A large proportion of potential respondents (~42%) exited the survey without completing a single item beyond consent, ostensibly because the required exertion of time and mental energy was not worth the return on investment.

5.3.3 Consequences of Sample Limitations

The inability of either round of field testing to produce an appropriate sample is considered a limiting factor of this research for the methodological modifications it prompted. First of all, there was the issue of the classical item analysis. Removal of the Masters' students from the Phase IIIA data reduced the usable sample down to a mere 35 respondents – a number far shy of the original target of 100. This prompted the decision to delay a thorough analysis of the item and instrument properties until after Phase IIIB data was in hand. Incidentally, this decision also eliminated a potential item-revision opportunity as the instrument had to remain intact across both phases if the data was ever to be merged.

The Phase IIIB recruitment resulted in only an additional 112 responses – again, a number far shy of the target of 500. Consequently, this prompted the investigation as to whether the data sets should/not be merged and if so, whether that would even raise the number of responses to a serviceable level. Essentially, instead of a local field test for instrument validation and calibration and then a national field test to collect data on *respondent* performance (as opposed to *instrument* performance), the study was reduced to a single field test of 16 institutions with an uneven and slightly inflated average cluster size owing to the disproportionate number of participants from VT. The same data then had to be used in both calibration and in testing, which is generally not the recommended practice.

The original goals of this dissertation were to build (and validate) an instrument measuring p-value fluency and then use that instrument to report baseline data on a national scale. The target number of 500 respondents was selected based on guidelines provided in methodological texts. For instance, de Ayala (2009) remarks that under favorable conditions, "it appears that a calibration sample of at least 500 persons and instruments of 20 or more items tend to produce reasonably accurate item parameter estimates" (p. 105) in the context of the 2PL model. For the 3PL model, samples of 1000 persons "may lead to reasonably accurate parameter estimates…under favorable conditions…however, it is strongly recommended that calibration sample sizes exceed 1000 to mitigate the convergence problems that sometimes plague 3PL model calibrations" (p. 130-131). While there is no consensus as to the precise number of participants needed, de Ayala suggests "a rough sample size guideline is that a calibration sample should have a few hundred respondents" (p. 43), subject to the proposed model, the length of the instrument, and the underlying dimensionality. The inability to reach even the least of these guidelines affected both goals of the dissertation.

Instrument validation was affected by sample size as discussed in Chapter 4. The inability to establish the dimensionality of the construct was a direct consequence of the instability in parameter estimates and diagnostics that accompany small samples. Furthermore, while item parameter estimates were calculated (albeit interpreted cautiously), person parameter estimates were not even attempted. Indirectly, the small sample size also limited the efforts at establishing norms for the target population as it is rather unrealistic to contend that a sample of 147 students representing only 16 institutions is sufficiently representative of all doctoral students nationally.

5.3.4 Summary of Limitations

In summary, this study suffered from both limitations, which affected the generalizability of the results, and limiting factors, which affected the methodological approach. Limitations included the homogeneity of the pilot study participants, the skewed nature of respondents' student status (i.e., year in school) in both the pilot study and the local field test, and unsatisfactory recruitment methods coupled with insufficient incentivization. The inability to reach the target sample size was both a limiting factor because of the many methodological modifications it prompted (e.g., inability to perform item analysis on local field test data, forced merging of local/national data, inability to have separate calibration and testing samples, inability to calculate person parameter estimates) and also a limitation because of the way it confines the interpretation of the results (e.g., instability of item parameter estimates, skepticism regarding the representativeness of the sample). These limitations, to the extent that it is possible, will be rectified in future rounds of data collection.

5.4 Future Directions

In many ways, this research has raised more questions than it has answered. This dissertation has essentially been a pilot study of sorts for a larger instrument development project. Areas that could stand to be improved, revisited, or explored have been either alluded to or explicitly mentioned throughout this discussion. These will be reiterated here with sufficient elaboration as is warranted. As this research was conducted in phases, next steps have been identified within that structure and will be presented as such in this section.

5.4.1 Future Directions for Item Development and Instrument Construction

Arguably the most pressing unanswered question lingering at the conclusion of this study is the establishment of the latent construct. The specification of the internal model and the construct map are at present speculative and incomplete. The goal of the research was to design an instrument to measure *p*-value fluency under the assumption that the misinterpretations on Greenland’s list collectively exhausted the content domain for this construct. In that regard, the study met the stated objective: an instrument was constructed and validated that measures the specified content; however, maybe ultimately this theoretical framework is inadequate.

As Wilson explained, the motivation for instrument design can come either from the proposed construct or the available items, neither is *more* appropriate than the other as a starting point. In this research, Greenland's list (items) served as the motivation around which to build the p-value fluency construct under the assumption that those who hold these misinterpretations would not be exhibiting fluency and would be unable to interpret and report *p*-values in the manner consistent with statistical theory and expert recommendation. Looking forward, this approach will be reimagined from the opposite perspective. *P*-value fluency as a construct is meant to represent a capability and thus perhaps it is more appropriate to use the ASA principles as the theoretical framework for that construct and then look for inspiration for items elsewhere (e.g. Greenland's list among others). Greenland's list of misinterpretations will still factor heavily in the items design for the instrument as their correspondence with the ASA principles can be justified (certainly for the 18 already in use, but the remaining 7 will also be considered), however, that will not be the sole basis for the content.

In analyzing and scoring the respondents' self-articulated *p*-value definitions, many (45/147) of these responses predictably expressed notions consistent with misinterpretations on Greenland's list (these are in red text for ease of identification, Appendix AB); however, this was not universally true. Among the remaining responses, multiple previously unidentified misconceptions were communicated calling into question the comprehensiveness of Greenland's list. As a point of clarification, this paper is not claiming that Greenland's list is incomplete for its intended purpose (a cataloging of *p*-value misinterpretations that have distorted the scientific literature), but rather that the list as a realization of the construct is insufficient. As the breadth of the content domain is investigated, a thorough analysis of these responses will be conducted with an eye to extraction of additional misinterpretations.

Inherent in defining the construct, beyond identifying the underlying list of concepts, is the determination of appropriate process levels. The present study was unable to resolve the uncertainty regarding the need for both contextual and context-free items as requisite to demonstration of *p*-value fluency. Data analysis (RQ2) was conducted on the validated subset of contextual items (vignettes) only, but the option was left open to revisit the inclusion of the context-free items (P/CP) in future iterations. The existing P/CP items were only partially validated and the next step is to improve the technical quality of those items with *negative* evidence. Beyond that, new items – both P/CP and Vignette style – will need to be written for any supplemental content that is included in the new internal model and construct map. This may involve expanding the vignette coverage to include alternate testing scenarios (ANOVA, Chi-square, etc.) or might simply mean adding the additional items to the existing scenarios that have already been well-vetted. Likely, both options will be investigated and item performance will designate the appropriate path forward.

The present study proposes two alternate test blueprints based on the current realization of the construct and the available validated items. The conclusion of the next phase of item writing will be to revise the test blueprint with an eye to organizing the instrument around 6 subscales (based on the ASA principles) and at two process levels (in/out of context) where there are at minimum, two items per subscale/process level combination. For reference, a blueprint constructed in this design with the set of available items would consist of 18 items distributed across 6 subscales at one process level (ASA1 = 3, ASA2 = 4, ASA3 = 1, ASA4 = 3, ASA5 = 3, ASA6 = 4). The reinsertion of the P/CP items would flesh out the second process level and the addition of newly identified misinterpretations would augment the number of items per subscale. Specifically, item development efforts would need to address ASA principle #3 as there is currently only one item for that subscale. Future research would revisit the idea of two concurrent blueprint options – both a full and partial version (or as they are typically referred, a *long* form for diagnosis and a *short* form for screening). The former would be more readily applicable to the context of a qualifying/preliminary exam or methods course exit assessment where the stakes are high and the test length is somewhat irrelevant; the latter option would be more appropriate for soliciting additional data from a broad audience when low investment and improper incentivization demand attention to such a detail for the preservation of response and completion rates.

5.4.2 Future Directions for Instrument Validation

Instrument validation is not so much a 'future direction' as it is the continuation of a work already in progress. The current condition of the instrument and validation efforts present two equally appropriate, but quite divergent paths forward. On the one hand, there is much to be gained from sticking with the current version of the instrument, *KPVMI-1*, and pursuing a larger, more nationally representative sample. If numbers could meet or exceed the recommended guidelines (e.g. at least 500 for 2PL models as explained in Section 5.3.3), the stable model parameters and person fit statistics could be computed. This would afford the opportunity to address those items on the Action Plan (Table 41) still pending.

The other option is to regard all the instrument development efforts thus far as exploratory analysis that will be used to inform a new instrument development project. In other words, the entire process will be revisited from the items design onward (item panel, cognitive interviews, pilot study, and field test will all be repeated). Assuming the item development goes as prescribed in 5.4.1, there will be supplemental items being added to the instrument in existing and novel dimensions. It would be unwise and shortsighted to presume that any prudent intermediate steps along the path to instrument development performed in the first trial could be abbreviated or eliminated in the second trial.

Wilson advises that "…any sound instrument is the result of several iterations, both part and whole, through [the steps[53]]" (2005, p.161). Adopting the first path describes a partial iteration through the steps; the latter path, a whole iteration. While the second option is certainly more labor intensive and time-consuming, there doesn't seem to be decent justification not to choose this option in light of the speculation regarding the definition of the construct and its underlying dimensionality. Adopting the ASA principles as the theoretical basis for the construct and the factor structure must be defended with empirical evidence and thus the full iteration of the instrument development process must be undertaken. The silver lining, if there is one, is that the process won't be initiated entirely *from scratch*. Any and all validity evidence gathered along the way necessarily informs future iterations of the process. The pool of validated items remains available for use and provides a base from which to re-initiate the cycle. Furthermore, even *negative* evidence has its usefulness. Lessons learned regarding incentivization and recruitment can and will be applied to future iterations which should improve the quality of future data collection efforts.

5.4.3 Future Directions for Data Analysis

The first two sections of this discussion addressed future directions that were directly related to the first research question. In this section, the second research question will be addressed. As it has been proposed that the instrument development cycle repeat, there will be a

[53] 'The steps' include the construct map, the items design, the outcome space, the measurement model, and quality control (reliability and validity).

preliminary need to repeat all the basic data analysis conducted in this dissertation with the new data. Beyond that, additional research will explore the subgroup results. Specifically, this study suggested the presence of an interaction effect between Major and Status, as well as between Major and Experience. Also suggested by this research was the notion that some ASA principles are universally elusive (across subgroups), but others fluctuate by Major. With an eye to applying these results at the curricular level, this is an important distinction to unpack. On an individual level, the correlation between subscales, as well as the relationship between knowledge of the p-value definition and performance on the instrument (in total and by subscale), will be further investigated beyond that which was presented in this research. Lastly, data on respondent training, opinion of its sufficiency, and initiative to exceed stated requirements was collected but only partially analyzed in this dissertation. Going forward, the relationships among these variables will be studied with an eye to measuring the accuracy of self-assessment and how that affects motivation.

5.5 Conclusions

The purpose of this study was to investigate on a national scale the baseline level of p-value fluency of future researchers (i.e., doctoral students). In pursuit of that objective, two research questions were investigated.

The first research question, *Can a sufficiently reliable and valid measure of p-value misinterpretations (in a research context) be constructed*?, has been deemed satisfactorily answered with the identification of a sufficiently valid and reliable collection of 30 items (12 from the P/CP section and 18 from the odd Vignette section). *PV3*, as it is referred, when considered on the whole, is a adequately reliable measure (*Alpha* = .8030) of *p*-value fluency as assessed across 18 misinterpretations (the entirety of the theoretical framework) and 2 process levels; when considered in part, it contains an independently validated sub-measure of *p*-value fluency *in context* as assessed across all 18 misinterpretations (*Alpha* = .8298).

The second research question, *What do the results of the KPVMI-1 administration tell us about the current level of p-value fluency among doctoral students nationally?,* was satisfactorily addressed (considering the sample limitations) with an analysis conducted on the Vignette subsection of items only (*KPVMI(P))*. This decision ensured that any implications drawn about the *p*-value fluency of doctoral students nationally was based on the most reliable subset of available items. In general, the median score was 10/18 items correct, or slightly better than half (56%). Small, but wholly insignificant, differences in mean performance were seen in the subgroup comparisons for Sex and Major.

This dissertation makes multiple contributions to statistics education research. First, there is the *KPVMI* instrument itself. The lack of an existing instrument to measure *p*-value fluency in a research context and suitable for use with the doctoral student population was identified upon review of the extant literature. The psychometric properties of the *KPVMI* are sufficient to tentatively offer this instrument as a reasonably viable mechanism for fulfilling that need. Secondly, evidence of doctoral students' (in)ability to identify misinterpretations and misuse of the *p*-value in practice was documented using a sufficiently reliable and valid instrument. This data augments what was previously known in this area based on research studies investigating practicing researchers and studies conducted with students in introductory courses. Finally, this research proposes the *p*-value fluency construct as a latent trait marrying a realization of the construct as manifested through research-identified misinterpretations with a subscale structure based on the recommended practice of the ASA's six principles for proper usage and reporting of *p*-values.

The goal of the research was to design an instrument to measure *p*-value fluency under the assumption that the misinterpretations on Greenland's list collectively exhausted the content domain for this construct. In that regard, the study met the stated objective: an instrument was constructed and validated that measured the specified content and generated meaningful data from a national administration. Limitations concerning the size and quality of the sample, coupled with the ambiguity regarding the underlying structure of the latent construct, prevent this research from being able to make broad generalizations regarding the level of *p*-value fluency on a national scale, but these preliminary results offer meaningful insight into the types of future directions this research should pursue in order to eventually achieve that objective.

References

Allen, K. (2006, May 1). *The Statistics Concept Inventory: Development and Analysis of a Cognitive Assessment Instrument in Statistics.* Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2130143

Allen, M.J. & Yen, W.M. (1979). *Introduction to Measurement Theory*. Brooks/Cole Publishers.

Babu, J. (2012). Bayesian and frequentist approaches. *Proceedings of the 7th Conference on Astronomical Data Analysis*.

Badenes-Ribera, L., Frias-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2015). Interpretation of the p-value: A national survey study in academic psychologists from Spain. *Psicothema*, *27*(3), 290-295.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 1-29.

Baker, M. (2017, August 15). Over half of psychology studies fail reproducibility test. *Nature*.

Beaty, B., & Torok, M. (2014). P-values: Democrats or dictators? *SAS Global Forum 2014 Proceedings*, (pp. 1-4).

Berger, J. (2003). Could Fisher, Jeffreys, and Neyman have agreed on testing? *Statistical Science, 18*(1), 1-32.

Box, J. (1978). *R.A. Fisher: The life of a scientist.* Wiley.

Bracey, G. (2008). How to avoid statistical traps. In J. Ballantine, & J. Spade (Eds.), *A sociological approach to education* (3rd ed.). Thousand Oaks, CA: Pine Forge Press.

Cai, L., Thissen, D., du Toit, S. H. C. (2016). IRTPRO for Windows, V 4.2 [Computer software]. Lincolnwood, IL: Scientific Software International.

Campbell, J. (1982). Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology, 67*, 691-700.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cohen, J. (1994). The Earth is round (p < .05). *American Psychologist, 49*(12), 997-1003.

Cronbach, L. J. (*1951*). Coefficient alpha and the internal structure of tests. Psychometrika, *16*, 297-334.

de Ayala, R.J. (2009). *The theory and practice of item response theory*. New York, NY: The Guildford Press.

delMas, R., Ooms, A., Garfield, J., & Chance, B. (2006). Assessing students' statistical reasoning. *Proceedings of the Seventh International Conference on Teaching Statistics.* Salvador de Bahia, Brazil.

Deubel, P. (2008, August 21). Education decisions: Where's the evidence and research base? *THE Journal*.

Dillman, D., Smyth, J., & Christian, L. (2009). *Internet, Mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Dozo, N. (2015, Feb 23). Nerisa@neri-peri. Retrieved from http://www.nature.com/news/psychology-journal-bans-p-values-1.17001

Editor, W. N. (2015, March 5). *Academic journal bans p-value significance test*. Retrieved from StatsLife: https://www.statslife.org.uk/news/2116-academic-journal-bans-p-value-significance-test

Editors, T. (2001). The value of p. *Epidemiology, 12*(3), 286.

Falk, R., & Greenbaum, C. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, 75-98.

Field, A. (2009). *Discovering statistics using SPSS.* London: SAGE Publications Ltd.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*, 309-368.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, *98*(1), 39-82.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological), 17*(1), 69-78.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1971 [1935]). *The Design of Experiments* (7th ed.). New York: Hafner Publishing Company.

Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 22-38.

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematics Thinking and Learning*, 99-125.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In *A handbook for data analysis in the behavioral sciences* (pp. 311-339). New Jersey: Erlbaum.

Gissane, C. (2012). The P value, do you know what it means? *Fysioterapeuten*, 106-106.

Goodman, S. (2008). A dirty dozen: Twelve p-avlue misconceptions. *Seminars in Hematology, 45*(3), 135-140.

Gregoire, T. (2001). Biometry in the 21st century: Whither statistical inference? *Proceedings of the conference on Forest Biometry and Information Science*, (pp. 1-15). Greenwich, London.

Hacking, I. (1965). *Logic of statistical inference.* Cambridge: Cambridge University Press.

Halpin, P., & Stam, H. (2006). Inductive inference or inductive behavior: Fisher and Neyman-Pearson approaches to statistical testing in psychological research (1940-1960). *The American Journal of Psychology, 119*(4), 625-653.

Irizarry, R., Peng, R., & Leek, J. (2014, February 14). *On the scalability of statistical procedures: why the p-value bashers just don't get it.* Retrieved from Simply Statistics: http://simplystatistics.org/2014/02/14/on-the-scalability-of-statistical-procedures-why-the-p-value-bashers-just-dont-get-it/

Karlsson, E. (2014, November). *Why p values and statistical significance are worthless in science*. Retrieved from Debunking Denialism: https://debunkingdenialism.com/2014/11/10/why-p-values-and-statistical-significance-are-worthless-in-science/

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical - significance tests are not. *Theory & Psychology,* 67-90.

Lane-Getaz, S. (2007, June). Development and validation of a research-based assessment: Reasoning about p-values and statistical significance. Doctoral Dissertation: University of Minnesota.

Lane-Getaz, S. (2013). Development of a reliable measure of students' inferential reasoning ability. *Statistics Education Research Journal, 12*(1), 20-47.

Lehmann, E. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88*(424), 1242-1249.

Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher and Neyman-Pearson. *British Journal for the Philosophy of Science, 57*, 69-91.

Liu, H. J. (1998). A cross-cultural study of sex differences in statistical reasoning for college students in Taiwan and the United States. Doctoral dissertation, University of Minnesota, Minneapolis.

Martin, N., Hughes, J, & Fugelsang, J. (2017). The roles of experience, gender, and individual differences in statistical reasoning. *Statistics Education Research Journal, 16*, 454 - 475.

Matthews, R. (2017, April). The ASA's p-value statement, one year on. *Significance*, 38-40.

Meehl, P. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103-115.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York, NY: American Council on Education and Macmillan.

Messick, S. (1992). Validity of test interpretation and use. In M. Alkin (Ed.), *Encyclopedia of Educational Research* (6th ed., pp. 1487-1495). New York, NY: MacMillan.

Meyer, J (2018). jMetrik, V 4.1.1 [computer software]. Charlottesville, VA: Psychomeasurement Systems, LLC.

Meyer, J. P. (2014). *Applied Measurement with jMetrik*. New York, NY: Routledge.

Misunderstandings of p-values. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Misunderstandings_of_p-values

Morrison, D., & Henkel, R. (Eds.). (1970). *The significance test controversy: A reader.* Chicago: Aldine.

Neyman, J. P. (1933a). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*, 289-337.

Neyman, J. P. (1933b). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society, 29*(4), 492-510.

Neyman, J., & Pearson, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part 1. *Biometrika, 20A*, 175-240.

Norstrom, F. (2015). Poor quality in the reporting and use of statistical methods in public health - the case of unemployment and health. *Archives of Public Health: The Official journal of the Belgian Public Health Association, 73*(56).

Nunnally, J.C. (1978) Psychometric theory. 2nd Edition, McGraw-Hill, New York.

Nuzzo, R. (2015, April 16). Scientists perturbed by loss of stat tools to sift research fudge from fact. *Scientific American*.

Ogburn, W. (1940). Statistical trends. *Journal of the American Statistical Association*, 252-260.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 50*(302), 157-175.

Perezgonzalez, J. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, 1-11.

Picho, K., & Artino, Jr., A. (2016). 7 deadly sins in educational research. *Journal of Graduate Medical Education, 8*(4), 483-487.

Prometric. (2017). *Internal Psychometric Guidelines for Classical Test Theory.* Retrieved from https://www.prometric.com/en-us/news-and-resources/reference-materials/pages/Internal-Psychometric-Guidelines-for-Classical-Test-Theory.aspx

Ranstam, J. (1996). A common misconception about p-values and its consequences. *Acta Orthopaedica Scandinavica*, 505-507.

Rao, C. R. (1992). R. A. Fisher: The founder of modern statistics. *Statistical Science, 7*(1), 34-48.

Reston, E. (2005). The statistical literacy assessment scale. *Paper presented at the 55th Session of the International Statistical Institute*, (pp. 1-6). Sydney, Australia.

Robinson, D., & Wainer, H. (2001). *On the past and future of null hypothesis significance testing.* Educational Testing Service, Statistics & Research Division, Princeton, NJ.

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*(10), 1276-1284.

Rothman, K. (1998). Writing for Epidemiology. *Epidemiology, 9*(3), 333-337.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm.* London: Chapman and Hall.

Rozeboom, W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Sarewitz, D. (2016, May 11). The pressure to publish pushes down quality. *Nature, 533*(7602).

Savage, L. (1976). On rereading R.A. Fisher. *The Annals of Statistics*, 441-500.

Schonlau, M., Fricker, R., & Elliott, M. (2002). *Conducting research surveys via e-mail and the web.* Santa Monica, CA: RAND Corporation.

Siegfried, T. (2010, March 27). Odds are, it's wrong. *ScienceNews*, p. 26.

Sotos, A., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, 98-113.

Spanos, A. (2013, February 17). *R.A. Fisher: How an outsider revolutionized statistics*. Retrieved from Error Statistics Philosophy: https://errorstatistics.com/2013/02/17/r-a-fisher-how-an-outsider-revolutionized-statistics/

Spiegelhalter, D. (2017, April). Too familiar to ditch. *Significance*, 41.

Suter, W. (2012). *Introduction to educational research: A critical thinking approach* (2nd ed.). Thousand Oaks, CA: Sage Publishing.

Tempelaar, D., Gijselaers, W., & Schim van der Loeff, S. (2006). Puzzles in statistical reasoning. *Journal of Statistics Education, 14*(1).

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1-2.

Vidgen, B, & Yasseri, T. (2016). P-values: Misunderstood and misused. *Frontiers in Physics*.

Warren, C., Denley, K., & Atchley, E. (2014). *Beginning Statistics* (2nd ed.). Mount Pleasant, SC: Hawkes Learning Systems/Quant Systems, Inc.

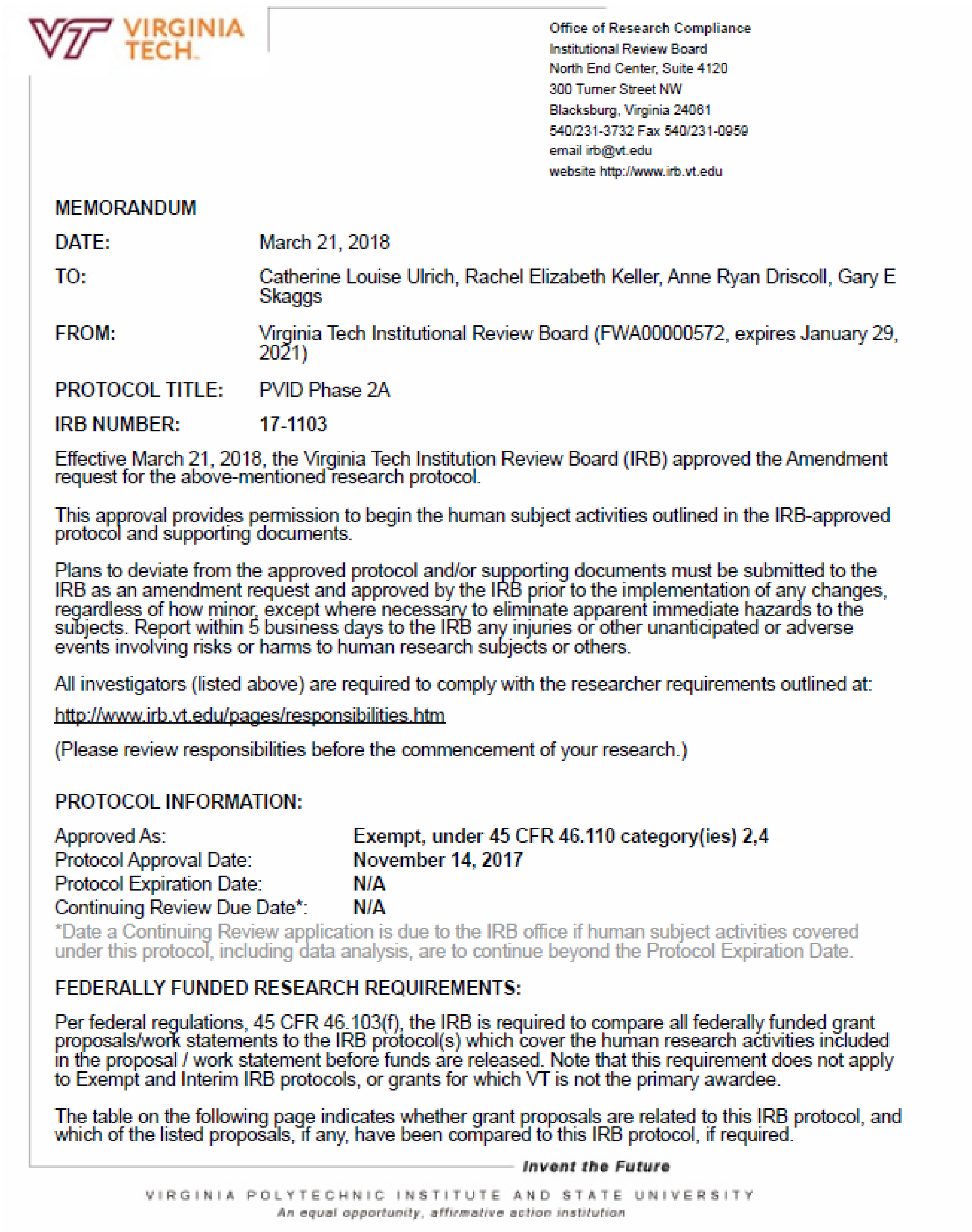
Wasserstein, R. (2016). American statistical association releases statement on statistical significance and p-values. *ASA News*, 1-3.

Wasserstein, R., & Lazar, N. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 1-17.

Wilson, M. (2005). *Constructing Measures: An item response modeling approach.* New York, NY: Psychology Press: Taylor & Francis Group.

Woolston, C. (2015). Psychology journal bans p values. *Nature, 519*(9).

Appendix A – IRB Documents for Phase IIA



VIRGINIA TECH.

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-3732 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE:	March 21, 2018
TO:	Catherine Louise Ulrich, Rachel Elizabeth Keller, Anne Ryan Driscoll, Gary E Skaggs
FROM:	Virginia Tech Institutional Review Board (FWA00000572, expires January 29, 2021)
PROTOCOL TITLE:	PVID Phase 2A
IRB NUMBER:	17-1103

Effective March 21, 2018, the Virginia Tech Institution Review Board (IRB) approved the Amendment request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Exempt, under 45 CFR 46.110 category(ies) 2,4**
Protocol Approval Date:	**November 14, 2017**
Protocol Expiration Date:	**N/A**
Continuing Review Due Date*:	**N/A**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

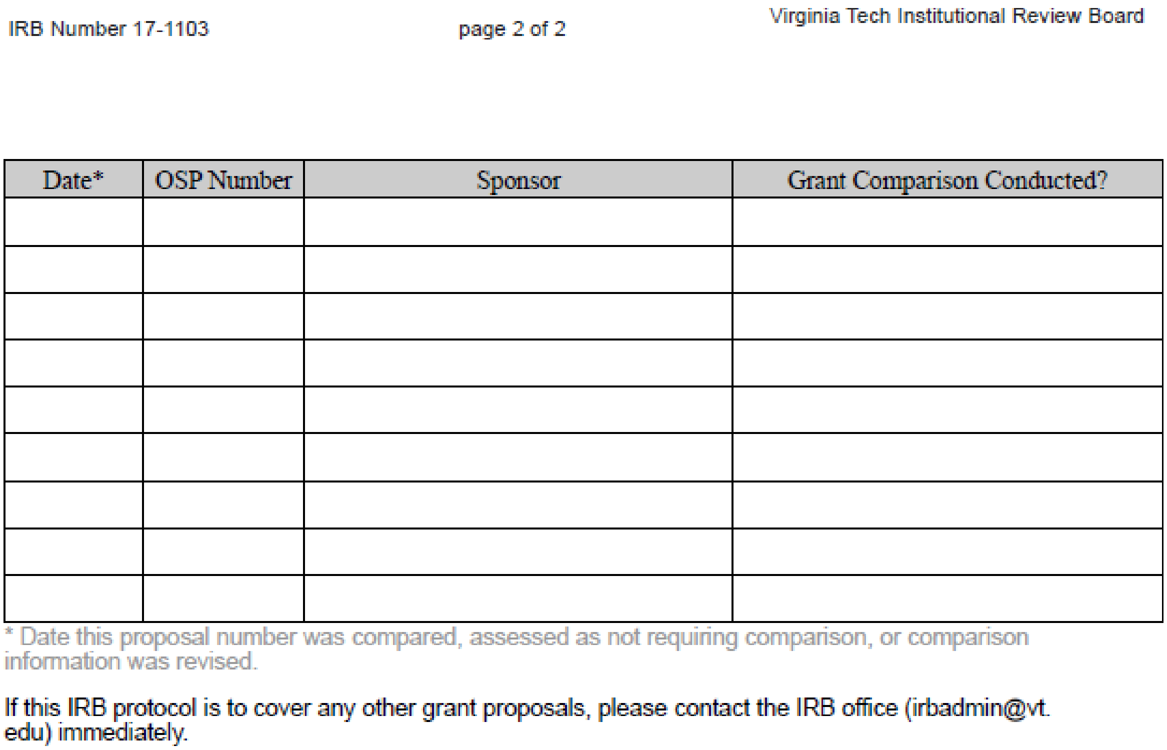
FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Invent the Future

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
An equal opportunity, affirmative action institution



IRB Number 17-1103 page 2 of 2 Virginia Tech Institutional Review Board

Date*	OSP Number	Sponsor	Grant Comparison Conducted?

* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this IRB protocol is to cover any other grant proposals, please contact the IRB office (irbadmin@vt.edu) immediately.

Recruitment Email

Hello,

I am conducting a research study for my PhD dissertation in Mathematics Education and I am inviting you to participate. For my project, I am interested in building an instrument to assess the extent to which doctoral students, i.e. future researchers, struggle with interpreting and reporting *p*-values in the context of independent research and peer review. Growing concern over the ability of practicing researchers to appropriately interpret and report statistical results has compelled journal editors and professional societies alike to take action in an effort to address and ameliorate the situation. In an effort to investigate whether the renewed attention to this issue has initiated a cycle of self-correction amongst the research community, the purpose of this research is to determine the extent to which *future* researchers are positioned to perpetuate this damaging pattern.

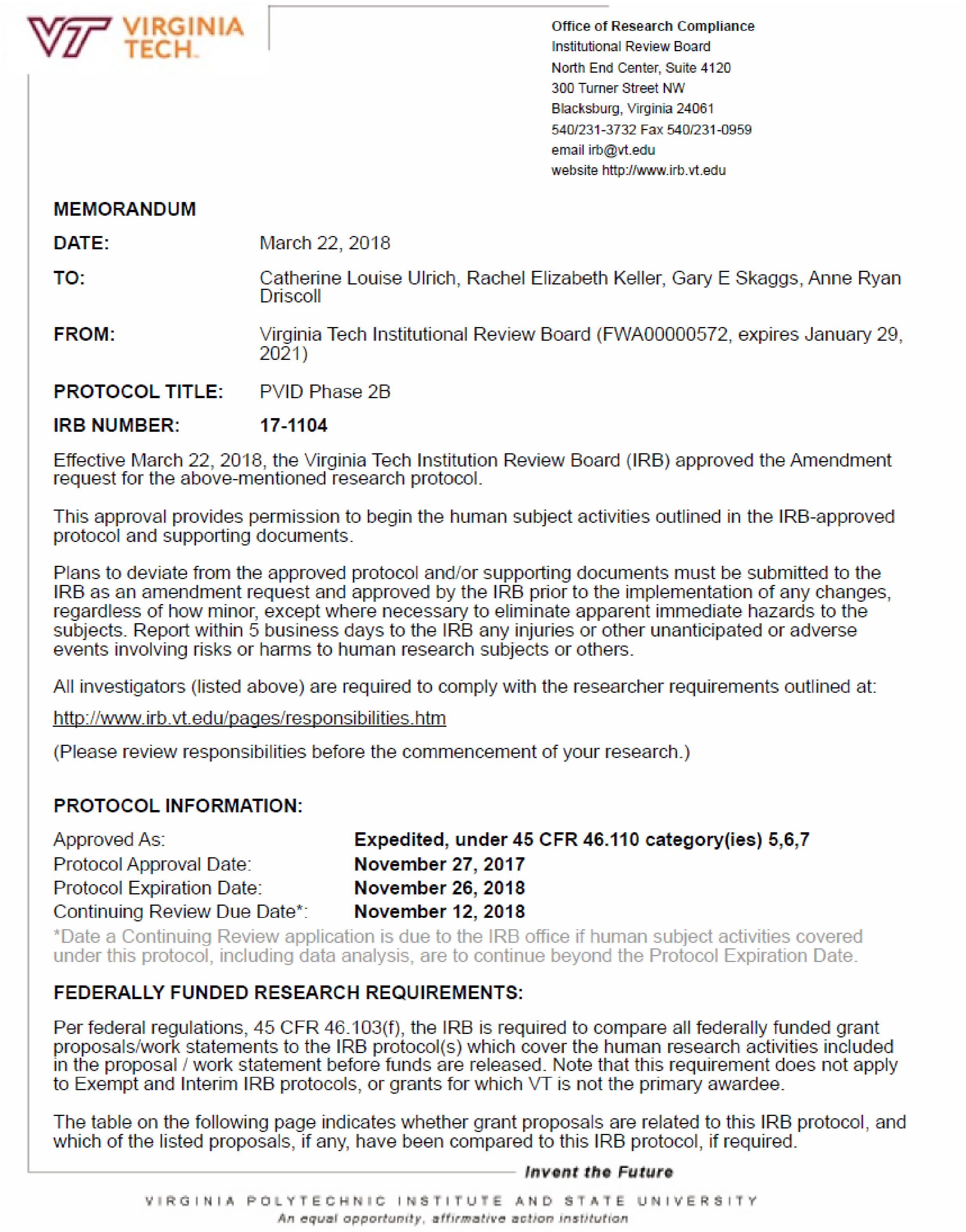
I am inviting you to participate in the pilot test for this instrument. The extent of your participation is that you complete the assessment and an exit questionnaire, for which you will be financially compensated for your time. This instrument will assess your ability to evaluate the research of others and to make decisions mimicking those that might arise in your own research. Part 1 of the assessment consists of 23 TRUE/FALSE items based on p-value misinterpretations. In Part 2 of the assessment, you will be asked to read research vignettes and evaluate whether the researcher's conclusions were valid or appropriate interpretations of *p*-values in context (36 statements in total distributed across 8 vignettes). The exit questionnaire will solicit your opinions regarding the usability of the instrument. The length of time for completion is different for everyone, but we estimate that it will require no more than 60 minutes. Please make every effort to complete all items.

Your participation is greatly appreciated; however, please note that participation is voluntary. You are free to withdraw from the study at any time or not complete all of the survey items. Your compensation will be a function of your level of participation: $5 upon completion of Part 1, $5 upon completion of Part 2, and $5 upon completion of the exit questionnaire. Should you have any questions about the study, you may contact me (rakeller@vt.edu) or Professor Ulrich (culrich@vt.edu). If you have concerns about the study's conduct or your rights as a research subject, or need to report a research-related injury or event, you may contact the VT Institutional Review Board (irb@vt.edu).

Thank you,

Rachel Keller

Appendix B – IRB Documents for Phase IIB



VIRGINIA TECH.

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-3732 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE:	March 22, 2018
TO:	Catherine Louise Ulrich, Rachel Elizabeth Keller, Gary E Skaggs, Anne Ryan Driscoll
FROM:	Virginia Tech Institutional Review Board (FWA00000572, expires January 29, 2021)
PROTOCOL TITLE:	PVID Phase 2B
IRB NUMBER:	**17-1104**

Effective March 22, 2018, the Virginia Tech Institution Review Board (IRB) approved the Amendment request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Expedited, under 45 CFR 46.110 category(ies) 5,6,7**
Protocol Approval Date:	**November 27, 2017**
Protocol Expiration Date:	**November 26, 2018**
Continuing Review Due Date*:	**November 12, 2018**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

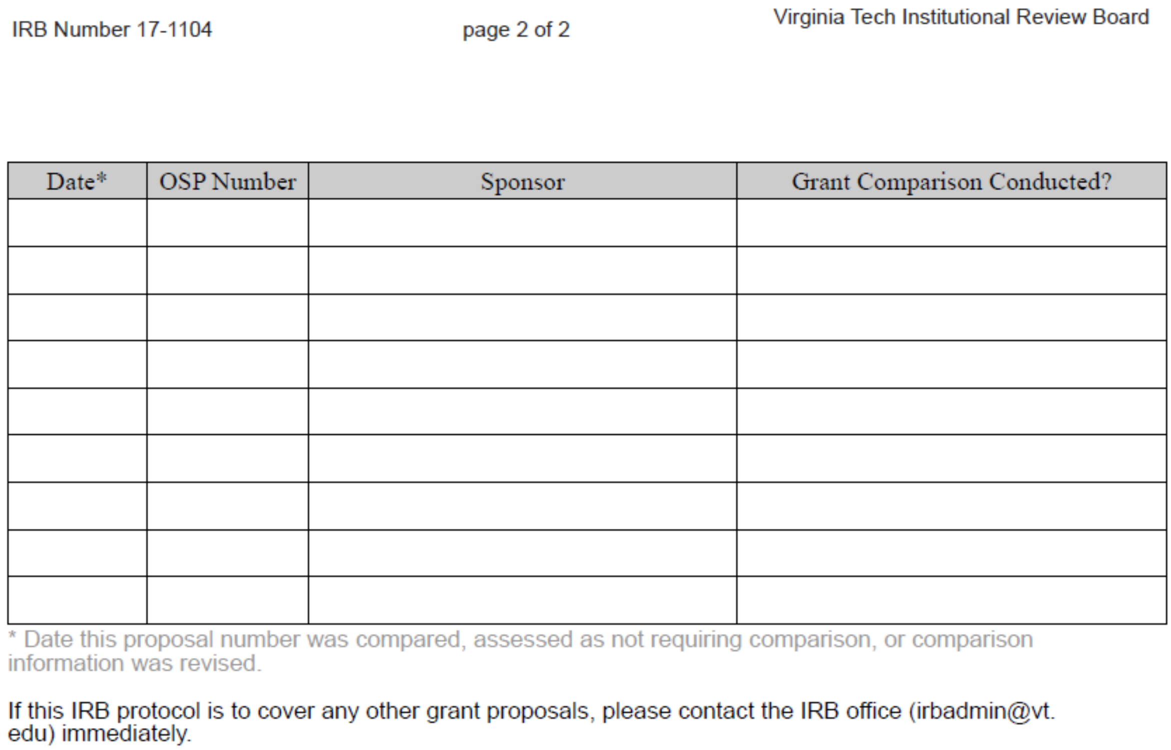
FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Invent the Future

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
An equal opportunity, affirmative action institution



Date*	OSP Number	Sponsor	Grant Comparison Conducted?

* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this IRB protocol is to cover any other grant proposals, please contact the IRB office (irbadmin@vt.edu) immediately.

Recruitment Email

Hello,

I am conducting a research study for my PhD dissertation in Mathematics Education and I am inviting you to participate. For my project, I am interested in building an instrument to assess the extent to which doctoral students, i.e. future researchers, struggle with interpreting and reporting *p*-values in the context of independent research and peer review. Growing concern over the ability of practicing researchers to appropriately interpret and report statistical results has compelled journal editors and professional societies alike to take action in an effort to address and ameliorate the situation. In an effort to investigate whether the renewed attention to this issue has initiated a cycle of self-correction amongst the research community, the purpose of this research is to determine the extent to which *future* researchers are positioned to perpetuate this damaging pattern.

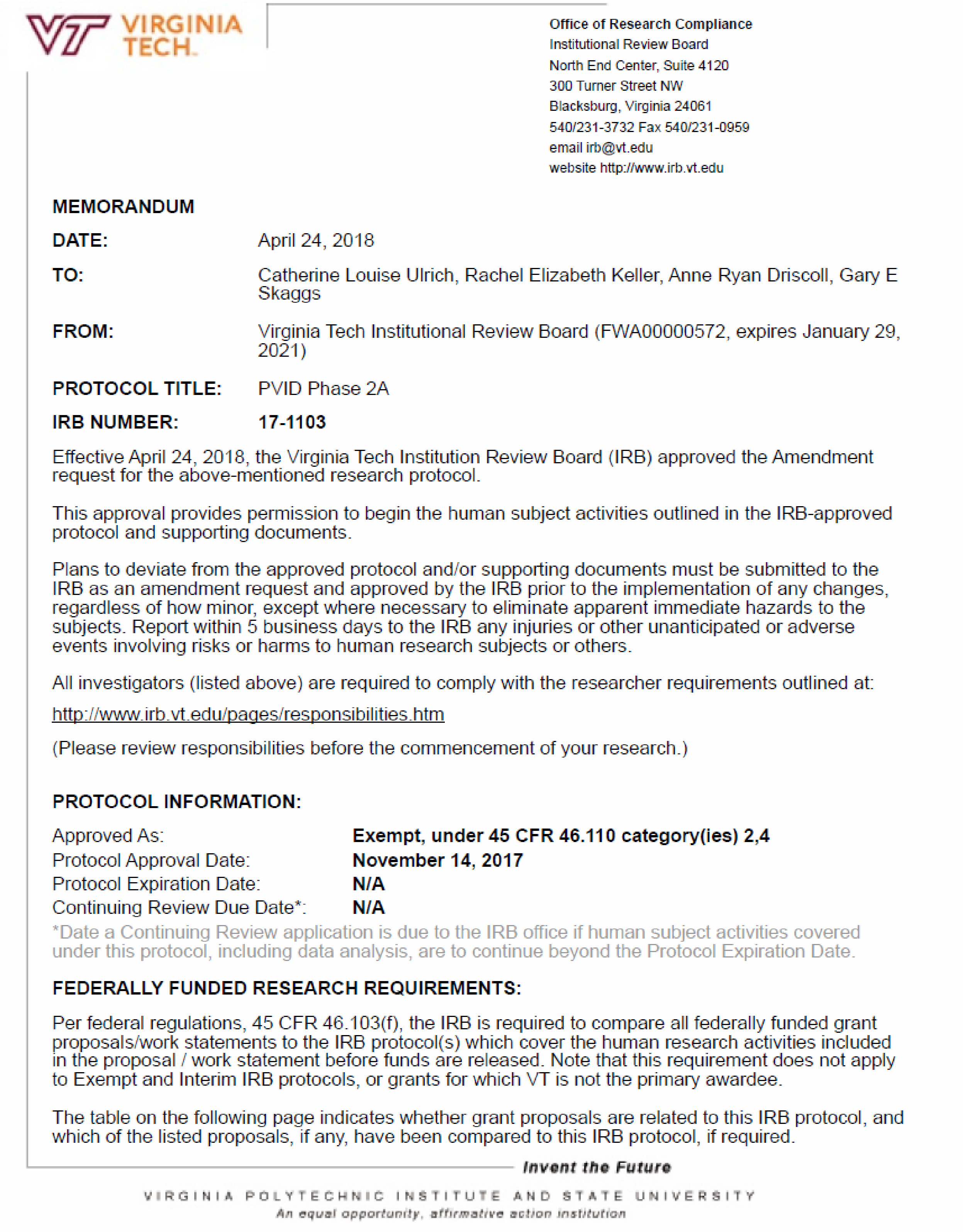
I am inviting you to participate in the pilot test for this instrument. The extent of your participation is that you complete the online assessment in the presence of the researcher, for which you will be financially compensated for your time at the rate of $15/hour. Participation time will vary by participant but we anticipate it will take between 1-2 hours. This instrument will assess your ability to evaluate the research of others and to make decisions mimicking those that might arise in your own research. Part 1 of the assessment consists of 23 TRUE/FALSE items based on p-value misinterpretations. In Part 2 of the assessment, you will be asked to read research vignettes and evaluate whether the researcher's conclusions were valid or appropriate interpretations of *p*-values in context (18 statements distributed across 4 vignettes). Upon completion of the assessment, you will be asked to comment on the overall usability of the instrument and make suggestions therein. Your activities will be video and audio recorded. You will be asked to "think aloud" as you answer the items, explaining your reasoning regarding your selections. You may also be asked to elaborate on items that require clarification for you to answer. The primary purpose of this phase of data collection is to obtain data regarding the usability of the assessment. The research team will use information from this study to modify the instrument in order to improve the study experience of future participants.

The consent form and research description is attached to this email. If you are interested in participating, please read that form and reply to this email indicating your desire to participate. Should you have any questions about the study, you may contact me (rakeller@vt.edu) or Professor Ulrich (culrich@vt.edu). If you have concerns about the study's conduct or your rights as a research subject, or need to report a research-related injury or event, you may contact the VT Institutional Review Board (irb@vt.edu).

Thank you,

Rachel Keller

Appendix C – IRB Documents for Phase IIIA



VIRGINIA TECH.

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-3732 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE:	April 24, 2018
TO:	Catherine Louise Ulrich, Rachel Elizabeth Keller, Anne Ryan Driscoll, Gary E Skaggs
FROM:	Virginia Tech Institutional Review Board (FWA00000572, expires January 29, 2021)
PROTOCOL TITLE:	PVID Phase 2A
IRB NUMBER:	**17-1103**

Effective April 24, 2018, the Virginia Tech Institution Review Board (IRB) approved the Amendment request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Exempt, under 45 CFR 46.110 category(ies) 2,4**
Protocol Approval Date:	**November 14, 2017**
Protocol Expiration Date:	**N/A**
Continuing Review Due Date*:	**N/A**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

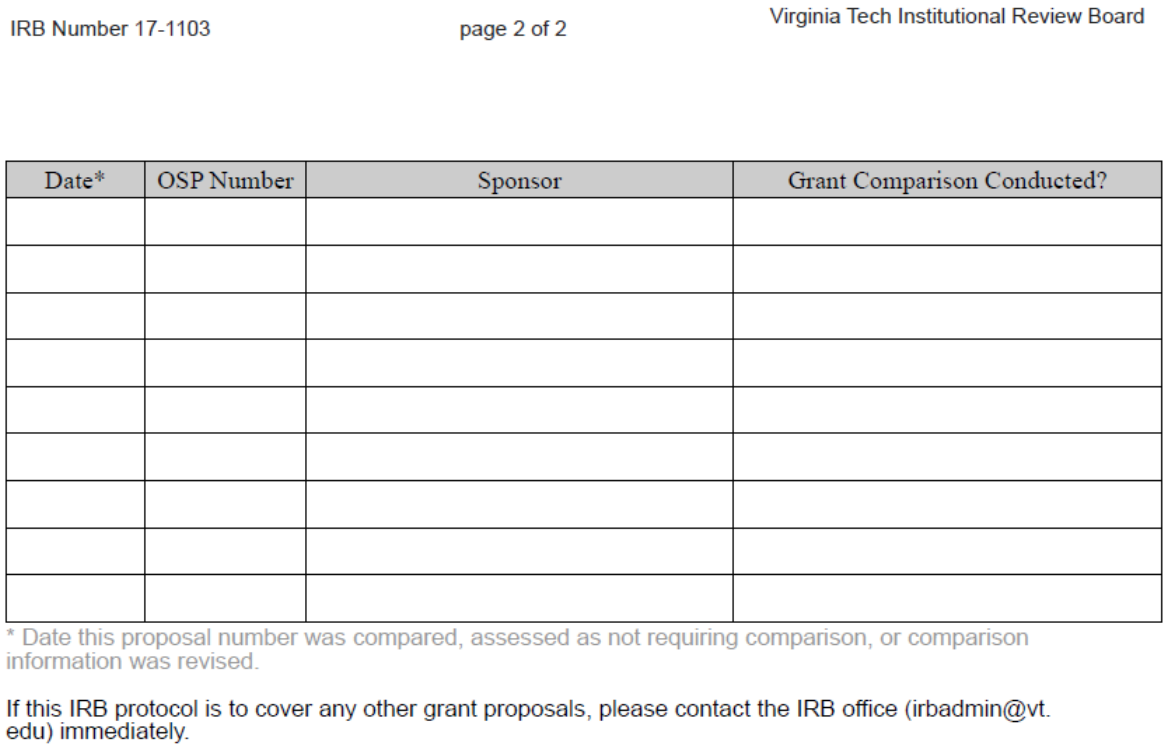
FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Invent the Future

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
An equal opportunity, affirmative action institution



IRB Number 17-1103 page 2 of 2 Virginia Tech Institutional Review Board

Date*	OSP Number	Sponsor	Grant Comparison Conducted?

* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this IRB protocol is to cover any other grant proposals, please contact the IRB office (irbadmin@vt.edu) immediately.

Recruitment Email

Hello,

I am conducting a research study for my PhD dissertation in Mathematics Education and I am inviting you to participate. For my project, I am interested in building an instrument to assess the extent to which doctoral students, i.e. future researchers, struggle with interpreting and reporting *p*-values in the context of independent research and peer review. Growing concern over the ability of practicing researchers to appropriately interpret and report statistical results has compelled journal editors and professional societies alike to take action in an effort to address and ameliorate the situation. In an effort to investigate whether the renewed attention to this issue has initiated a cycle of self-correction amongst the research community, the purpose of this research is to determine the extent to which *future* researchers are positioned to perpetuate this damaging pattern.

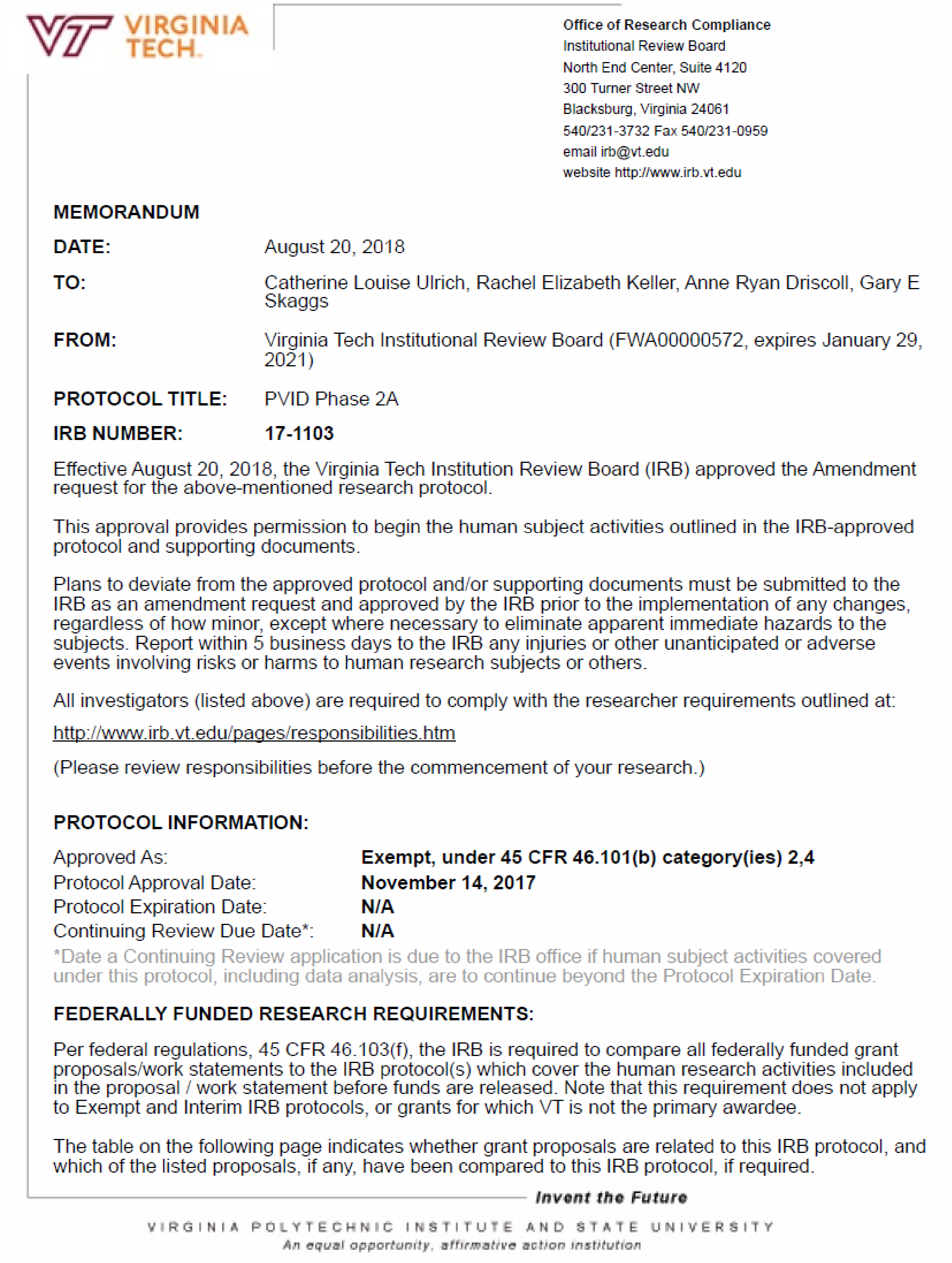
I am inviting you to participate in the field test for this instrument. The extent of your participation is that you complete the assessment and an exit questionnaire, for which you will be compensated for your time. This instrument will assess your ability to evaluate the research of others and to make decisions mimicking those that might arise in your own research. Part 1 of the assessment consists of TRUE/FALSE items based on p-value misinterpretations. In Part 2 of the assessment, you will be asked to read research vignettes and evaluate whether the researcher's actions were valid or appropriate interpretations of *p*-values in context. The exit questionnaire will solicit your opinions regarding the usability of the instrument and will collect demographic data. The length of time for completion is different for everyone, but we estimate that it will require no more than 60 minutes. Please make every effort to complete all items.

Your participation is greatly appreciated; however, please note that participation is voluntary. You are free to withdraw from the study at any time or not complete all of the survey items. Your compensation will be a function of your level of participation: 3 extra credit points will be awarded to your exam grade upon completion of all assessment items (demographic information may be omitted without penalty). Should you have any questions about the study, you may contact me (rakeller@vt.edu) or Professor Driscoll (adriscoll@vt.edu). If you have concerns about the study's conduct or your rights as a research subject, or need to report a research-related injury or event, you may contact the VT Institutional Review Board (irb@vt.edu).

Thank you,

Rachel Keller

Appendix D – IRB Documents for Phase IIIB



VIRGINIA TECH.

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-3732 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE:	August 20, 2018
TO:	Catherine Louise Ulrich, Rachel Elizabeth Keller, Anne Ryan Driscoll, Gary E Skaggs
FROM:	Virginia Tech Institutional Review Board (FWA00000572, expires January 29, 2021)
PROTOCOL TITLE:	PVID Phase 2A
IRB NUMBER:	**17-1103**

Effective August 20, 2018, the Virginia Tech Institution Review Board (IRB) approved the Amendment request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Exempt, under 45 CFR 46.101(b) category(ies) 2,4**
Protocol Approval Date:	**November 14, 2017**
Protocol Expiration Date:	**N/A**
Continuing Review Due Date*:	**N/A**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

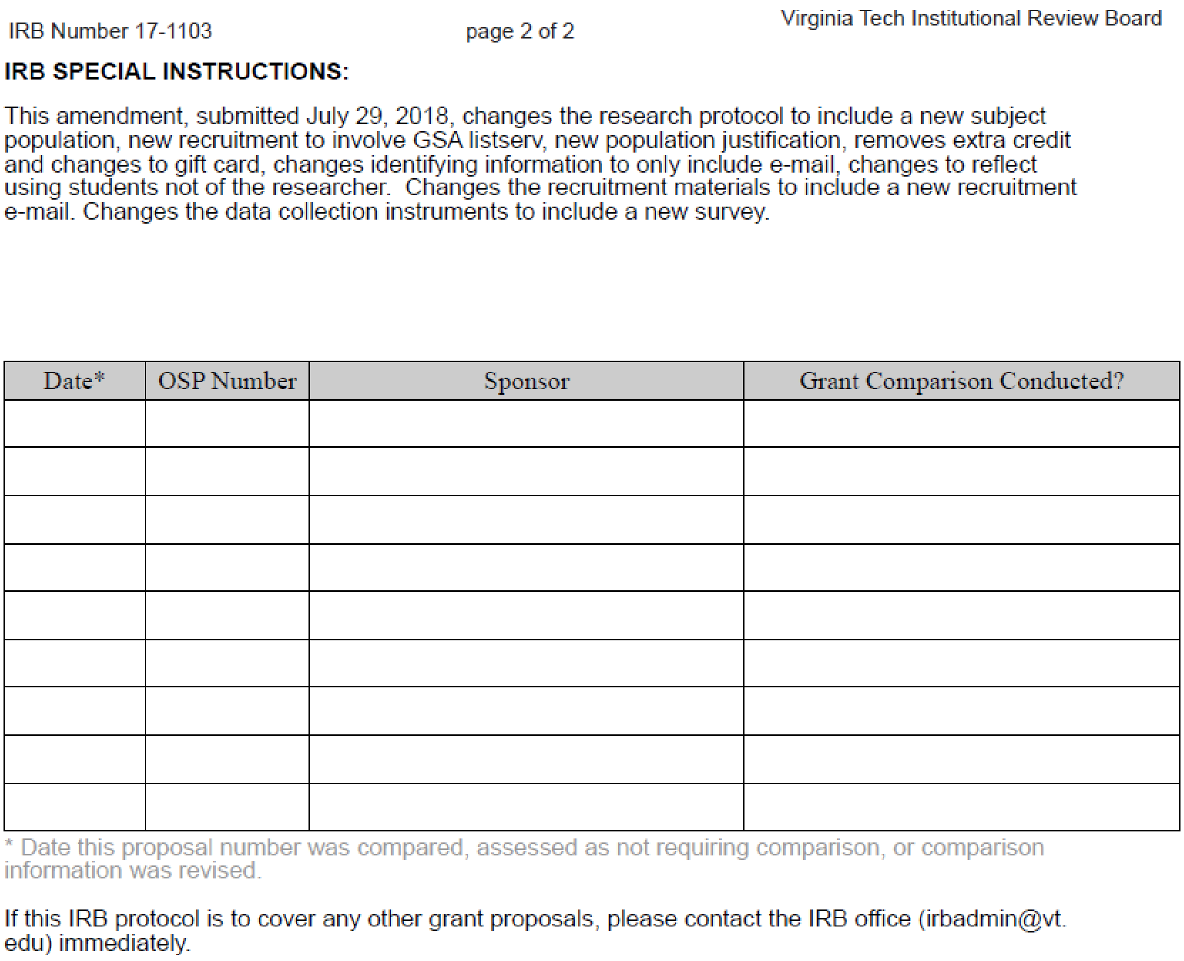
FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Invent the Future

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
An equal opportunity, affirmative action institution



IRB Number 17-1103 page 2 of 2 Virginia Tech Institutional Review Board

IRB SPECIAL INSTRUCTIONS:

This amendment, submitted July 29, 2018, changes the research protocol to include a new subject population, new recruitment to involve GSA listserv, new population justification, removes extra credit and changes to gift card, changes identifying information to only include e-mail, changes to reflect using students not of the researcher. Changes the recruitment materials to include a new recruitment e-mail. Changes the data collection instruments to include a new survey.

Date*	OSP Number	Sponsor	Grant Comparison Conducted?

* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this IRB protocol is to cover any other grant proposals, please contact the IRB office (irbadmin@vt.edu) immediately.

Recruitment Email for Graduate Student Organizations

Subject Line: Virginia Tech research study seeks your help in recruiting participants for national sample

Hello,

I am a PhD candidate at Virginia Polytechnic Institute & State University and I am asking for your support in my research. For my dissertation, I am interested in building an instrument to assess the extent to which doctoral students, i.e. future researchers, struggle with interpreting and reporting *p*-values in the context of independent research and peer review.

In an effort to collect a national sample of R1 institutions, I am inviting the doctoral students at your institution to participate in the field test for this instrument. Participation includes completion of the assessment and an exit questionnaire. I am asking for your assistance in advertising my study on your campus. At Virginia Tech, we have a graduate student listserve that generates weekly emails regarding research participation opportunities (along with events and services). It is my assumption that a similar or equivalent means of electronic communication exists at your institution (be it email or website or discussion forum, etc.) and I am asking if you would be willing to disseminate my recruitment materials via the delivery method you deem most appropriate.

I have prepared a recruitment letter (attached) that describes the research in greater detail and I have also furnished a short blurb (below) that can be used in the advertisement. Feel free to use either/both of these materials as you see fit. If you would like to know more about this research, please don't hesitate to contact me.

Thank you for your consideration.
Rachel Keller
rakeller@vt.edu
717-571-0000

Study Advertisement
National study conducted by Virginia Tech of p-value fluency in doctoral candidates is looking for participants. Participants will be asked to complete a two-part assessment (no more than 60 minutes) that will assess the ability to evaluate the research of others and to make decisions mimicking those that might arise in your own research. Part 1 of the assessment consists of 18 TRUE/FALSE items based on p-value misinterpretations. In Part 2 of the assessment, you will be asked to read 4 research vignettes and evaluate whether the researcher's actions in each scenario were valid or appropriate interpretations of *p*-values in context. Demographic information will also be collected. As an incentive for your participation, one participant at each institution, via raffle, will receive a digital giftcard of $10 to the merchant of their choice (Starbucks, Amazon, iTunes). Additionally, all participants will be entered into a larger drawing for one of seven digital gift cards valued at $50 (data collection will be terminated after 1 month or when 1150 completed responses have been recorded; probability of winning the large drawing are then no worse than 7/1150)**.** Participation is both voluntary and confidential. Interested persons can contact the researcher (rakeller@vt.edu) directly or access the instrument online at the following link: VTPMI (or https://virginiatech.qualtrics.com/jfe/form/SV_a9K2AQXLQgMIKsl)

Recruitment Email for Participants

Subject line: Virginia Tech research study invites you to be part of a national sample of PhD students

Hello,

I am conducting a research study for my PhD dissertation in Mathematics Education at Virginia Tech and I am inviting you to participate. For my project, I am interested in building an instrument to assess the extent to which doctoral students, i.e. future researchers, struggle with interpreting and reporting *p*-values in the context of independent research and peer review. Growing concern over the ability of practicing researchers to appropriately interpret and report statistical results has compelled journal editors and professional societies alike to take action in an effort to address and ameliorate the situation. In an effort to investigate whether the renewed attention to this issue has initiated a cycle of self-correction amongst the research community, the purpose of this research is to determine the extent to which *future* researchers are positioned to perpetuate this damaging pattern.

I am inviting you to participate in the field test for this instrument. The extent of your participation is that you complete the assessment and an exit questionnaire. This instrument will assess your ability to evaluate the research of others and to make decisions mimicking those that might arise in your own research. Part 1 of the assessment consists of TRUE/FALSE items based on p-value misinterpretations. In Part 2 of the assessment, you will be asked to read research vignettes and evaluate whether the researcher's actions were valid or appropriate interpretations of *p*-values in context. The exit questionnaire will solicit your opinions regarding the usability of the instrument and will collect demographic data. The length of time for completion is different for everyone, but we estimate that it will require no more than 60 minutes. Please make every effort to complete all items.

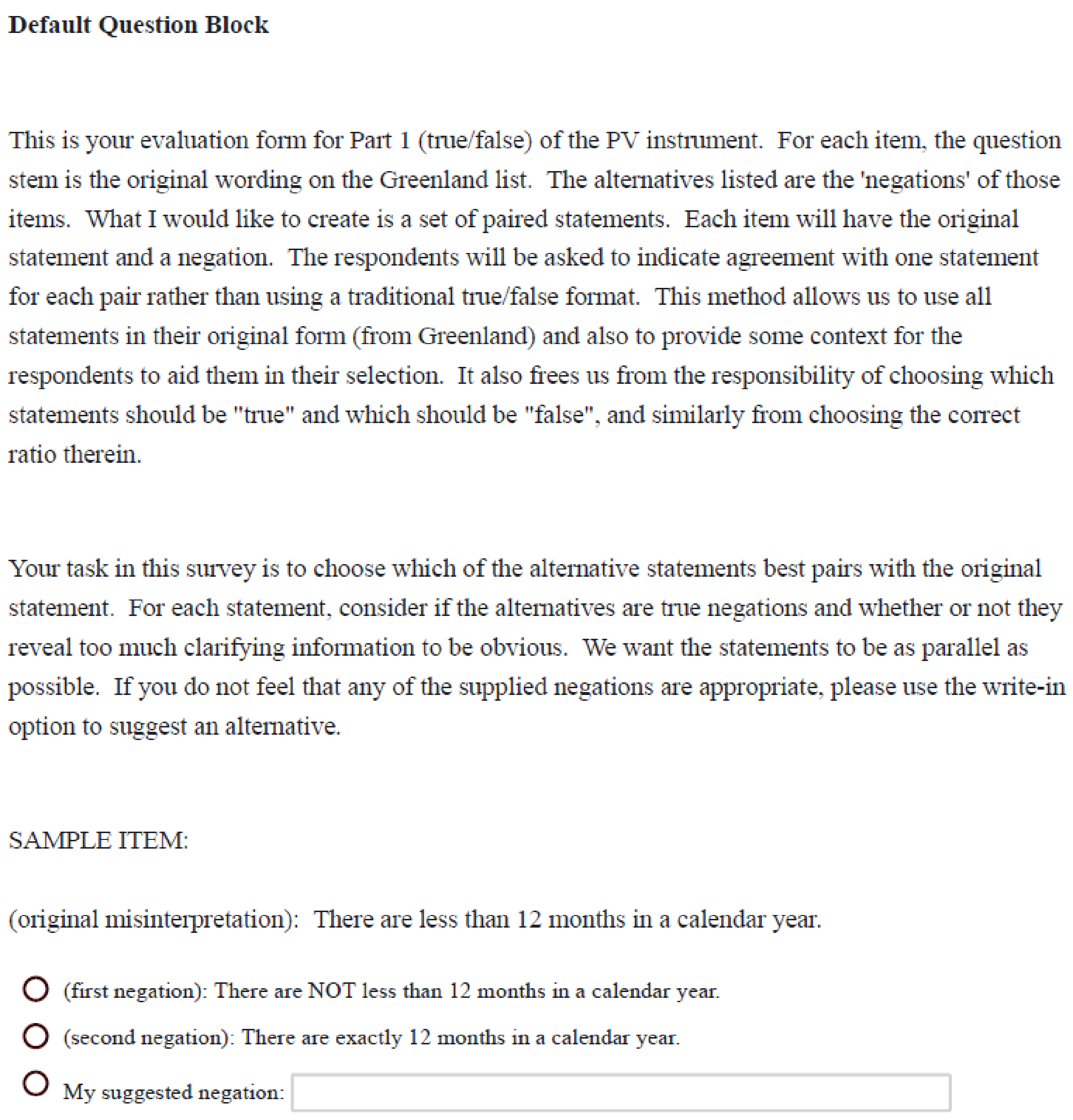
Your participation is greatly appreciated; however, please note that participation is voluntary. You are free to withdraw from the study at any time or not complete all of the survey items. This survey is being advertised at R1 institutions nationwide. As an incentive for your participation, one participant at each institution, via raffle, will receive a digital giftcard of $10 to the merchant of their choice (Starbucks, Amazon, iTunes). Additionally, all participants will be entered into a larger drawing for one of seven digital gift cards valued at $50 (data collection will be terminated after 1 month or when 1150 completed responses have been recorded; probability of winning the large drawing are then no worse than 7/1150; probability of winning is 7/n where n is the number of responses not to exceed 1150)**.** Participation is confidential. Should you have any questions about the study, you may contact me (rakeller@vt.edu) or Professor Driscoll (adriscoll@vt.edu). If you have concerns about the study's conduct or your rights as a research subject, or need to report a research-related injury or event, you may contact the VT Institutional Review Board (irb@vt.edu).

Thank you,

Rachel Keller

Link to Qualtrics: VTPMI (or https://virginiatech.qualtrics.com/jfe/form/SV_a9K2AQXLQgMIKsl)

Appendix E – PV0(1) Survey



Default Question Block

This is your evaluation form for Part 1 (true/false) of the PV instrument. For each item, the question stem is the original wording on the Greenland list. The alternatives listed are the 'negations' of those items. What I would like to create is a set of paired statements. Each item will have the original statement and a negation. The respondents will be asked to indicate agreement with one statement for each pair rather than using a traditional true/false format. This method allows us to use all statements in their original form (from Greenland) and also to provide some context for the respondents to aid them in their selection. It also frees us from the responsibility of choosing which statements should be "true" and which should be "false", and similarly from choosing the correct ratio therein.

Your task in this survey is to choose which of the alternative statements best pairs with the original statement. For each statement, consider if the alternatives are true negations and whether or not they reveal too much clarifying information to be obvious. We want the statements to be as parallel as possible. If you do not feel that any of the supplied negations are appropriate, please use the write-in option to suggest an alternative.

SAMPLE ITEM:

(original misinterpretation): There are less than 12 months in a calendar year.

○ (first negation): There are NOT less than 12 months in a calendar year.

○ (second negation): There are exactly 12 months in a calendar year.

○ My suggested negation:

What is your name?

Block 1

#1. The p-value is the probability that the test hypothesis is true.

○ The p-value is NOT a measure of the probability of the truth of the test hypothesis.

○ The p-value is NOT the probability that the test hypothesis is true.

○ The p-value is NOT a hypothesis probability for any hypothesis.

○ My suggested negation:

#2. The p-value for the null hypothesis is the probability that chance alone produced the observed association.

○ The probability that chance alone produced the observed association is NOT measured by the p-value.

○ The probability that chance alone produced the observed association canNOT be determined from the p-value.

○ The p-value is a probability computed assuming chance was operating alone and thus canNOT be the probability that chance produced the observed association.

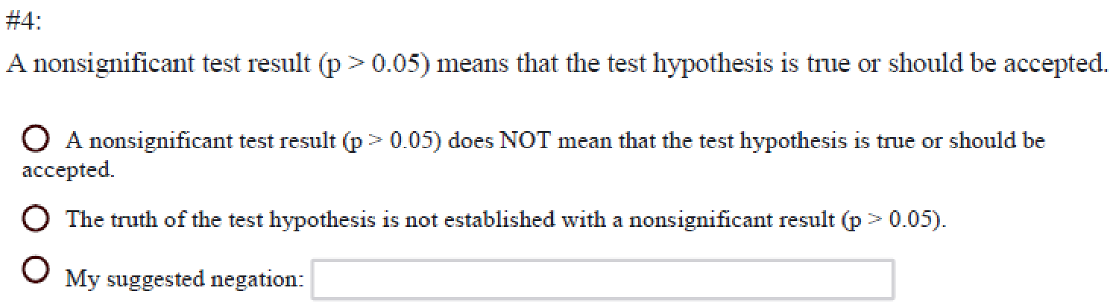
○ My suggested negation:

#3. A significant test result ($p \leq 0.05$) means that the test hypothesis is false or should be rejected.

○ A significant test result ($p \leq 0.05$) does NOT mean that the test hypothesis is false or should be rejected.

○ The falsity of the test hypothesis is not established with a significant result ($p \leq 0.05$).

○ My suggested negation:



#4:

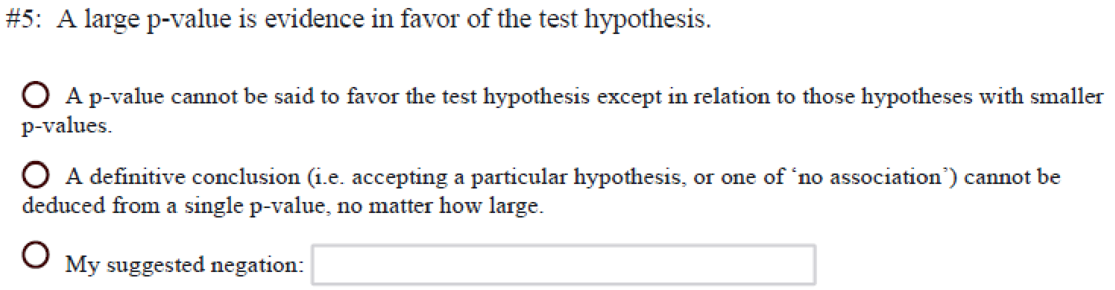
A nonsignificant test result (p > 0.05) means that the test hypothesis is true or should be accepted.

○ A nonsignificant test result (p > 0.05) does NOT mean that the test hypothesis is true or should be accepted.

○ The truth of the test hypothesis is not established with a nonsignificant result (p > 0.05).

○ My suggested negation:

Block 2

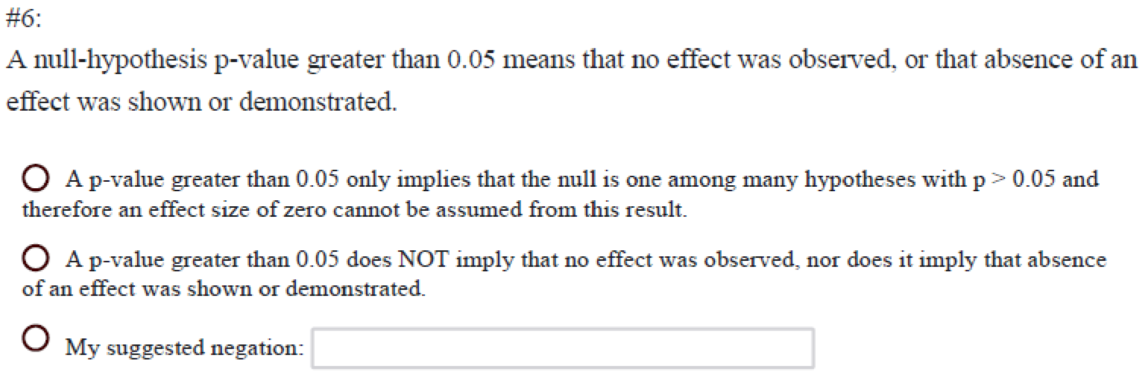


#5: A large p-value is evidence in favor of the test hypothesis.

○ A p-value cannot be said to favor the test hypothesis except in relation to those hypotheses with smaller p-values.

○ A definitive conclusion (i.e. accepting a particular hypothesis, or one of ‘no association’) cannot be deduced from a single p-value, no matter how large.

○ My suggested negation:



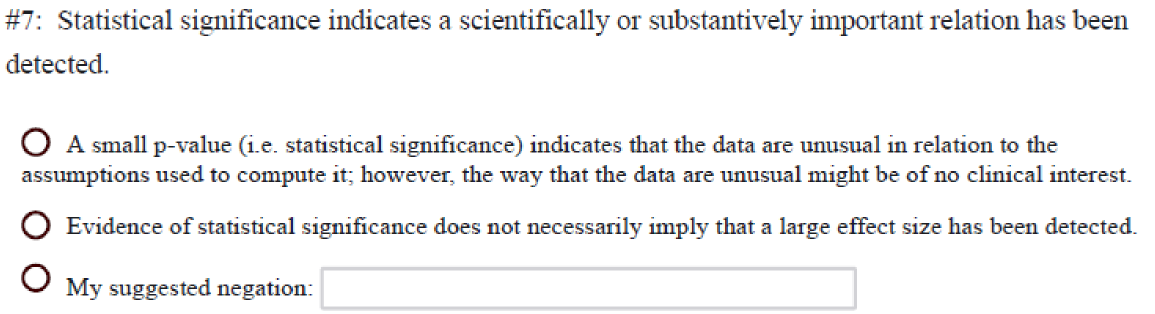
#6:

A null-hypothesis p-value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.

○ A p-value greater than 0.05 only implies that the null is one among many hypotheses with p > 0.05 and therefore an effect size of zero cannot be assumed from this result.

○ A p-value greater than 0.05 does NOT imply that no effect was observed, nor does it imply that absence of an effect was shown or demonstrated.

○ My suggested negation:

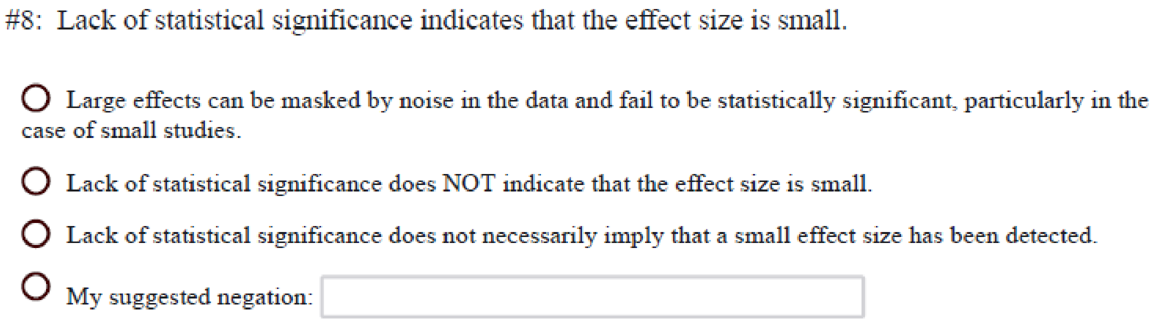


#7: Statistical significance indicates a scientifically or substantively important relation has been detected.

○ A small p-value (i.e. statistical significance) indicates that the data are unusual in relation to the assumptions used to compute it; however, the way that the data are unusual might be of no clinical interest.

○ Evidence of statistical significance does not necessarily imply that a large effect size has been detected.

○ My suggested negation:



#8: Lack of statistical significance indicates that the effect size is small.

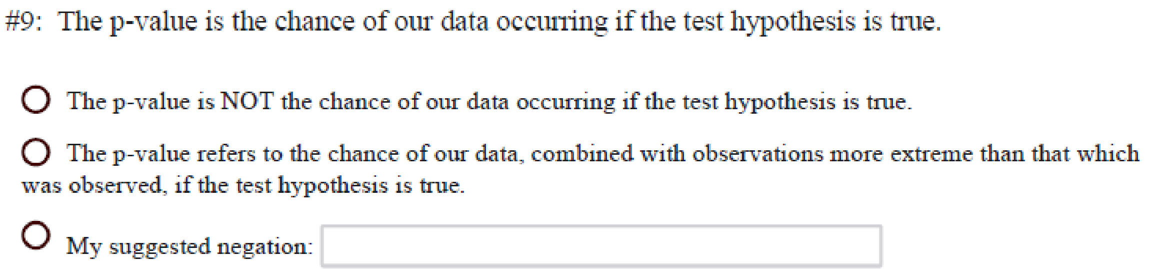
○ Large effects can be masked by noise in the data and fail to be statistically significant, particularly in the case of small studies.

○ Lack of statistical significance does NOT indicate that the effect size is small.

○ Lack of statistical significance does not necessarily imply that a small effect size has been detected.

○ My suggested negation:

Block 3

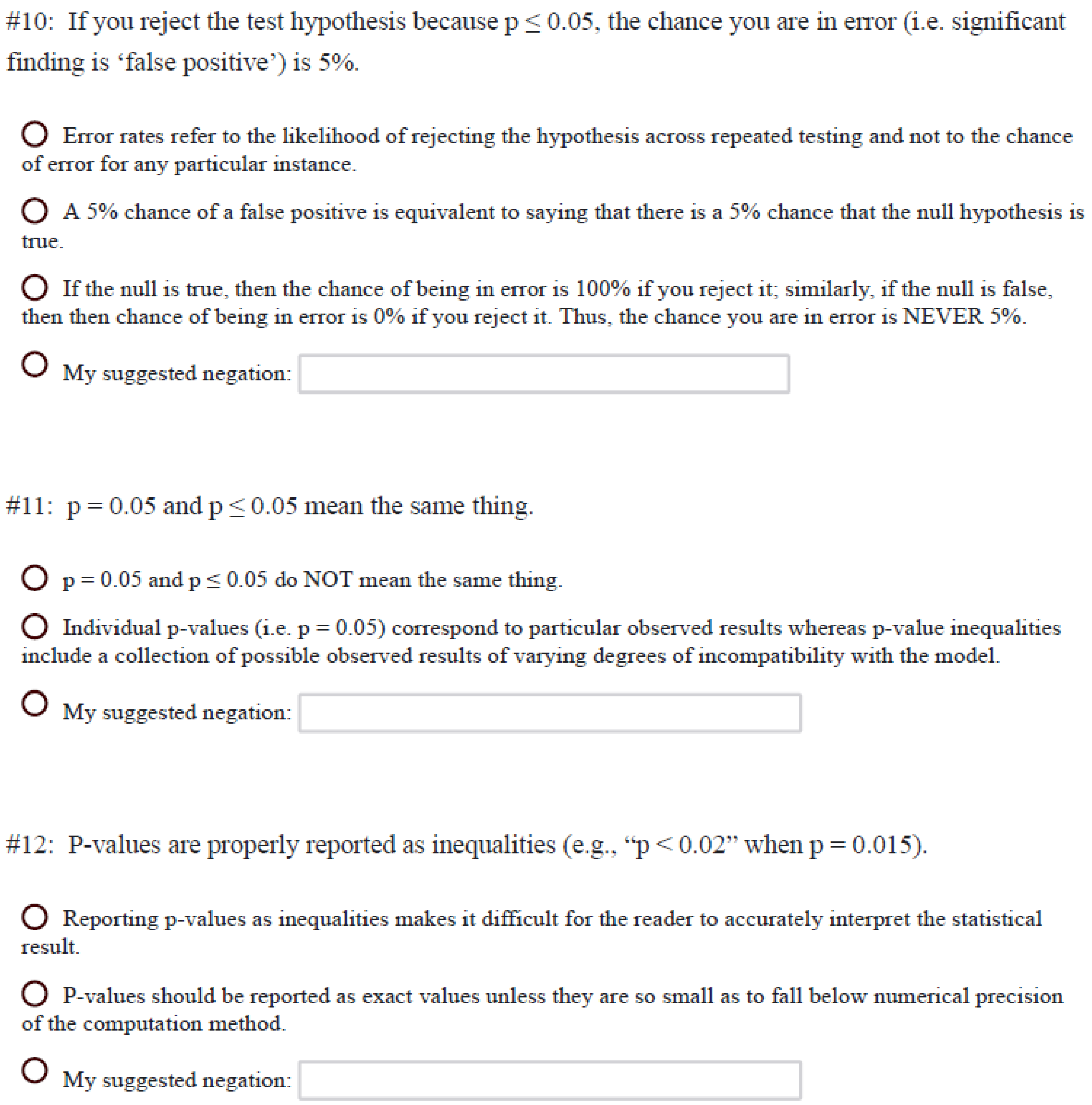


#9: The p-value is the chance of our data occurring if the test hypothesis is true.

○ The p-value is NOT the chance of our data occurring if the test hypothesis is true.

○ The p-value refers to the chance of our data, combined with observations more extreme than that which was observed, if the test hypothesis is true.

○ My suggested negation:



#10: If you reject the test hypothesis because $p \leq 0.05$, the chance you are in error (i.e. significant finding is ‘false positive’) is 5%.

○ Error rates refer to the likelihood of rejecting the hypothesis across repeated testing and not to the chance of error for any particular instance.

○ A 5% chance of a false positive is equivalent to saying that there is a 5% chance that the null hypothesis is true.

○ If the null is true, then the chance of being in error is 100% if you reject it; similarly, if the null is false, then then chance of being in error is 0% if you reject it. Thus, the chance you are in error is NEVER 5%.

○ My suggested negation:

#11: $p = 0.05$ and $p \leq 0.05$ mean the same thing.

○ $p = 0.05$ and $p \leq 0.05$ do NOT mean the same thing.

○ Individual p-values (i.e. $p = 0.05$) correspond to particular observed results whereas p-value inequalities include a collection of possible observed results of varying degrees of incompatibility with the model.

○ My suggested negation:

#12: P-values are properly reported as inequalities (e.g., “$p < 0.02$” when $p = 0.015$).

○ Reporting p-values as inequalities makes it difficult for the reader to accurately interpret the statistical result.

○ P-values should be reported as exact values unless they are so small as to fall below numerical precision of the computation method.

○ My suggested negation:

Block 4

#13: Statistical significance is a property of the phenomenon being studied and thus statistical tests detect significance.

○ Statistical significance is NOT a property of the phenomenon being studied and thus statistical tests do NOT detect significance.

○ Statistical significance is a property attached to the result of a statistical test and NOT a property of the effect or population being studied.

○ My suggested negation:

#14: One should always use two-sided p-values.

○ The nature of the p-value should be compatible with the wording of the hypotheses: one-sided hypotheses call for one-sided p-values, two-sided hypotheses call for two-sided p-values.

○ One-sided p-values are permitted under certain conditions.

○ Two-sided p-values can only be used under certain conditions.

○ My suggested negation:

#15: When the same hypothesis is tested in different studies and none (or a majority) of the tests are statistically significant (all $p > 0.05$), the overall evidence supports the hypothesis.

○ Individual studies could fail to reach statistical significance, but when combined could show persuasive evidence of an effect.

○ Lack of statistical significance of individual studies should not be taken as implying that the totality of evidence supports no effect.

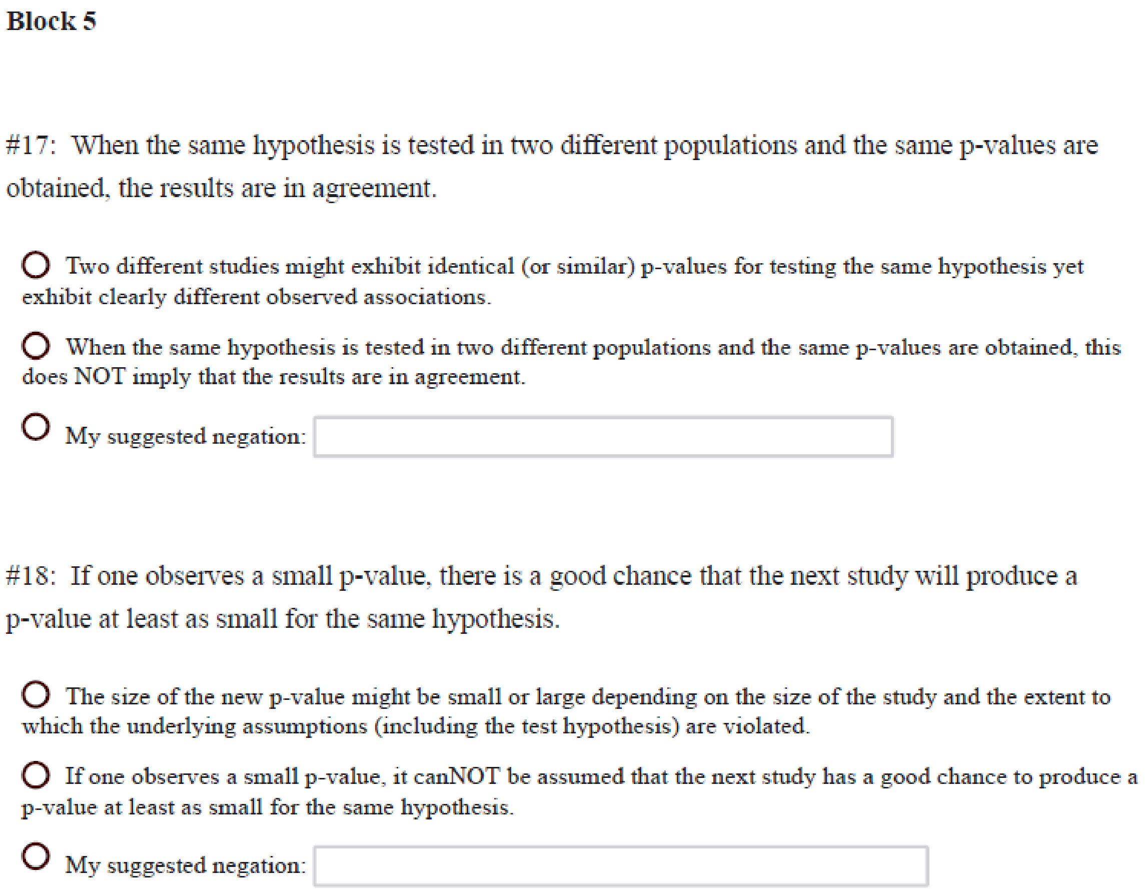
○ My suggested negation:

#16: When the same hypothesis is tested in two different populations and the resulting p-values are on opposite sides of 0.05, the results are conflicting.

○ Two studies may provide very different p-values for the same test hypothesis despite being in perfect agreement (e.g. may show identical observed associations).

○ When the same hypothesis is tested in two different populations and the resulting p-values are on opposite sides of 0.05, this does NOT imply that the results are conflicting.

○ My suggested negation:



Block 5

#17: When the same hypothesis is tested in two different populations and the same p-values are obtained, the results are in agreement.

○ Two different studies might exhibit identical (or similar) p-values for testing the same hypothesis yet exhibit clearly different observed associations.

○ When the same hypothesis is tested in two different populations and the same p-values are obtained, this does NOT imply that the results are in agreement.

○ My suggested negation:

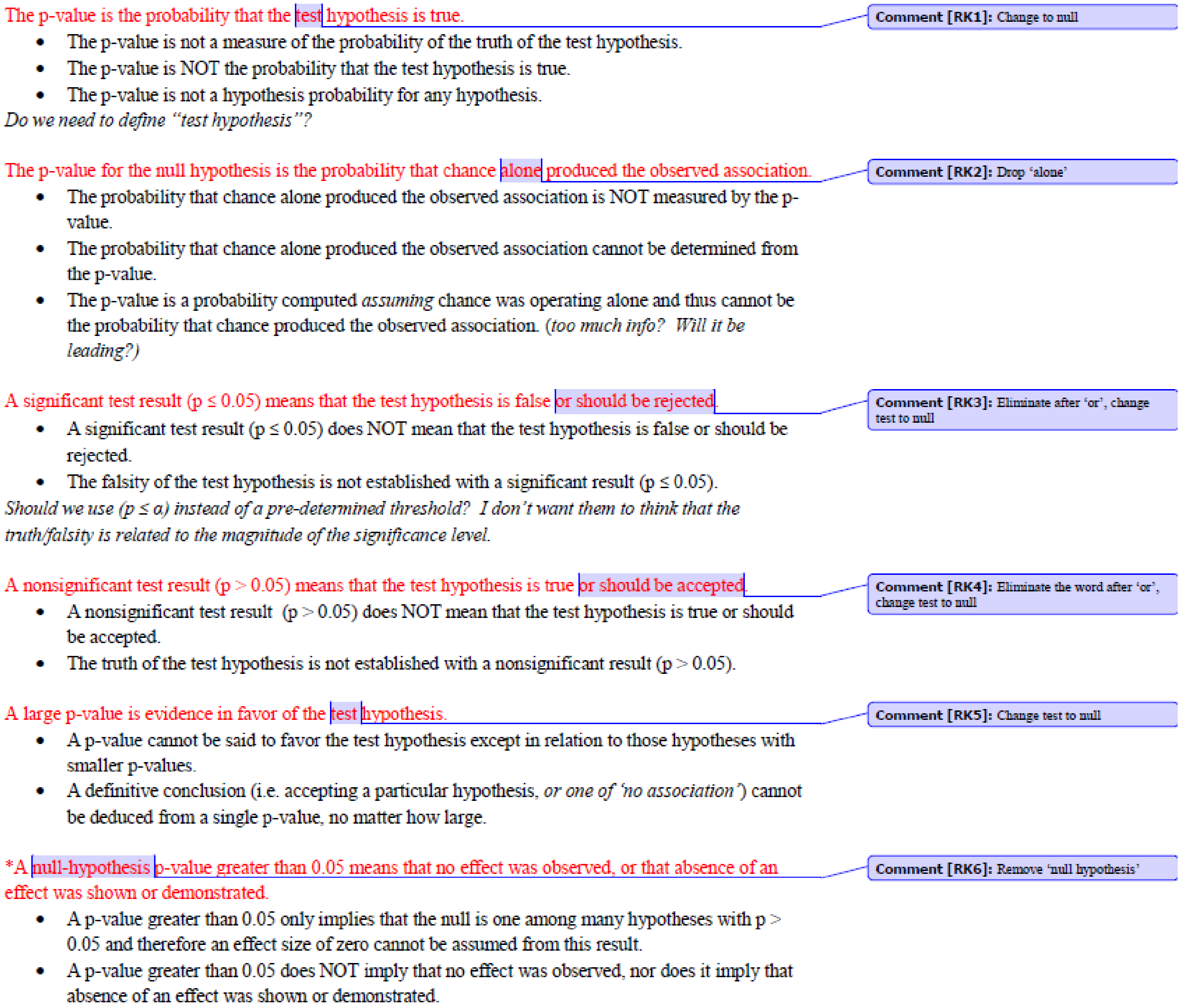
#18: If one observes a small p-value, there is a good chance that the next study will produce a p-value at least as small for the same hypothesis.

○ The size of the new p-value might be small or large depending on the size of the study and the extent to which the underlying assumptions (including the test hypothesis) are violated.

○ If one observes a small p-value, it canNOT be assumed that the next study has a good chance to produce a p-value at least as small for the same hypothesis.

○ My suggested negation:

Appendix F – Language Edits to Initial Item Pool



The p-value is the probability that the test hypothesis is true.

Comment [RK1]: Change to null

- The p-value is not a measure of the probability of the truth of the test hypothesis.
- The p-value is NOT the probability that the test hypothesis is true.
- The p-value is not a hypothesis probability for any hypothesis.

Do we need to define "test hypothesis"?

The p-value for the null hypothesis is the probability that chance alone produced the observed association.

Comment [RK2]: Drop 'alone'

- The probability that chance alone produced the observed association is NOT measured by the p-value.
- The probability that chance alone produced the observed association cannot be determined from the p-value.
- The p-value is a probability computed *assuming* chance was operating alone and thus cannot be the probability that chance produced the observed association. *(too much info? Will it be leading?)*

A significant test result (p ≤ 0.05) means that the test hypothesis is false or should be rejected.

Comment [RK3]: Eliminate after 'or', change test to null

- A significant test result (p ≤ 0.05) does NOT mean that the test hypothesis is false or should be rejected.
- The falsity of the test hypothesis is not established with a significant result (p ≤ 0.05).

Should we use (p ≤ α) instead of a pre-determined threshold? I don't want them to think that the truth/falsity is related to the magnitude of the significance level.

A nonsignificant test result (p > 0.05) means that the test hypothesis is true or should be accepted.

Comment [RK4]: Eliminate the word after 'or', change test to null

- A nonsignificant test result (p > 0.05) does NOT mean that the test hypothesis is true or should be accepted.
- The truth of the test hypothesis is not established with a nonsignificant result (p > 0.05).

A large p-value is evidence in favor of the test hypothesis.

Comment [RK5]: Change test to null

- A p-value cannot be said to favor the test hypothesis except in relation to those hypotheses with smaller p-values.
- A definitive conclusion (i.e. accepting a particular hypothesis, *or one of 'no association'*) cannot be deduced from a single p-value, no matter how large.

*A null-hypothesis p-value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.

Comment [RK6]: Remove 'null hypothesis'

- A p-value greater than 0.05 only implies that the null is one among many hypotheses with p > 0.05 and therefore an effect size of zero cannot be assumed from this result.
- A p-value greater than 0.05 does NOT imply that no effect was observed, nor does it imply that absence of an effect was shown or demonstrated.

Statistical significance indicates a scientifically or substantively important relation has been detected.

- A small p-value (i.e. statistical significance) indicates that the data are unusual in relation to the assumptions used to compute it; however, the way that the data are unusual might be of no clinical interest.
- Evidence of statistical significance does not necessarily imply that a large effect size has been detected.
- *Item about large sample sizes?*

Lack of statistical significance indicates that the effect size is small.

- Large effects can be masked by noise in the data and fail to be statistically significant, particularly in the case of small studies.
- Lack of statistical significance does NOT indicate that the effect size is small.
- Lack of statistical significance does not necessarily imply that a small effect size has been detected.

*The p-value is the chance of our data occurring if the test hypothesis is true.

Comment [RK7]: Change test to null

- The p-value refers to the chance of our data, combined with observations *more extreme* than that which was observed, if the test hypothesis is true.
- The p-value is NOT the chance of our data occurring if the test hypothesis is true.

If you reject the test hypothesis because $p \le 0.05$, the chance you are in error (i.e. significant finding is ‘false positive’) is 5%.

Comment [RK8]: Change test to null, the chance you have committed a type I error

- Error rates refer to the likelihood of rejecting the hypothesis across repeated testing and not to the chance of error for any particular instance.
- A 5% chance of a false positive is equivalent to saying that there is a 5% chance that the null hypothesis is true
- If the null is true, then the chance of being in error is 100% if you reject it; similarly, if the null is false, then then chance of being in error is 0% if you reject it. Thus, the chance you are in error is NEVER 5%.

*$P = 0.05$ and $p \le 0.05$ mean the same thing.

- $P = 0.05$ and $p \le 0.05$ do NOT mean the same thing.
- Individual p-values (i.e. $p = 0.05$) correspond to particular observed results whereas p-value inequalities include a collection of possible observed results of varying degrees of incompatibility with the model.

P-values are properly reported as inequalities (e.g., “$p < 0.02$” when $p = 0.015$).

- Reporting p-values as inequalities makes it difficult for the reader to accurately interpret the statistical result.
- P-values should be reported as exact values unless they are so small as to fall below numerical precision of the computation method.

Statistical significance is a property of the phenomenon being studied, and thus statistical tests detect significance.

Comment [RK9]: Cut after here

- Statistical significance is NOT a property of the phenomenon being studied and thus statistical tests do NOT detect significance.
- Statistical significance is a property attached to the result of a statistical test and not a property of the effect or population being studied.

*One should always use two-sided p-values.

- The nature of the p-value should be compatible with the wording of the hypotheses: one-sided hypotheses call for one-sided p-values, two-sided hypotheses call for two-sided p-values.
- Two-sided p-values should always be used, regardless of the nature of the hypotheses. (not a negation)
- One-sided p-values are permitted under certain conditions.
- Two-sided p-values can only be used under certain conditions.

When the same hypothesis is tested in different studies and none (or a majority) of the tests are statistically significant (all $p > 0.05$), the overall evidence supports the hypothesis.

Comment [RK10]: Add 'null' here

Comment [RK11]: minority

Comment [RK12]: Add the word 'null'

- Individual studies could fail to reach statistical significance, but when combined could show persuasive evidence of an effect.
- Lack of statistical significance of individual studies should not be taken as implying that the totality of evidence supports no effect.

When the same hypothesis is tested in two different populations and the resulting p-values are on opposite sides of 0.05, the results are conflicting.

Comment [RK13]: Change to 'twice'

- Two studies may provide very different p-values for the same test hypothesis despite being in perfect agreement (e.g. may show identical observed associations).
- When the same hypothesis is tested in two different populations and the resulting p-values are on opposite sides of 0.05, this does NOT imply that the results are conflicting.

When the same hypothesis is tested in two different populations and the same p-values are obtained, the results are in agreement.

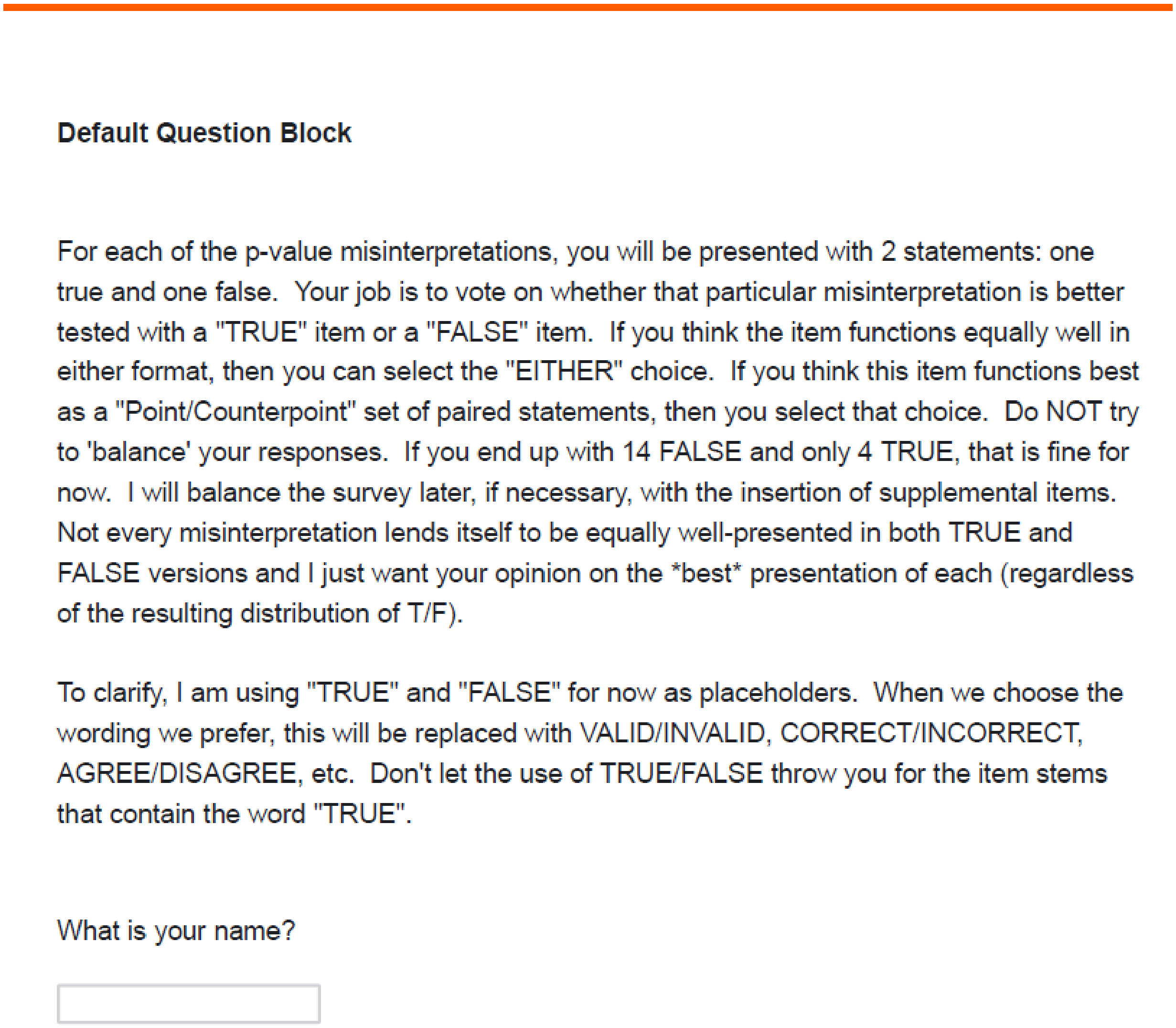
Comment [RK14]: twice

- Two different studies might exhibit identical (or similar) p-values for testing the same hypothesis yet exhibit clearly different observed associations.
- When the same hypothesis is tested in two different populations and the same p-values are obtained, this does NOT imply that the results are in agreement.

If one observes a small p-value, there is a good chance that the next study will produce a p-value at least as small for the same hypothesis.

- The size of the new p-value might be small or large depending on the size of the study and the extent to which the underlying assumptions (including the test hypothesis) are violated.
- If one observes a small p-value, it canNOT be assumed that the next study has a good chance to produce a p-value at least as small for the same hypothesis.

Appendix G – PV0(2) Survey

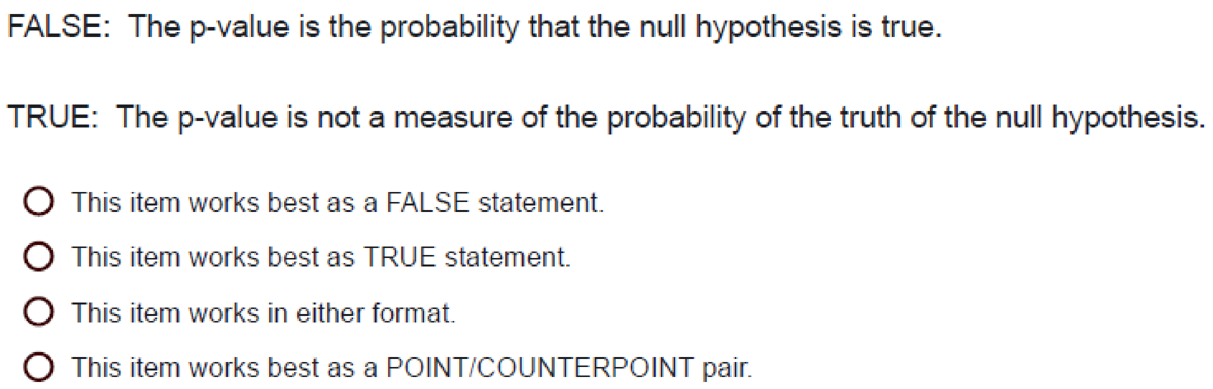


Default Question Block

For each of the p-value misinterpretations, you will be presented with 2 statements: one true and one false. Your job is to vote on whether that particular misinterpretation is better tested with a "TRUE" item or a "FALSE" item. If you think the item functions equally well in either format, then you can select the "EITHER" choice. If you think this item functions best as a "Point/Counterpoint" set of paired statements, then you select that choice. Do NOT try to 'balance' your responses. If you end up with 14 FALSE and only 4 TRUE, that is fine for now. I will balance the survey later, if necessary, with the insertion of supplemental items. Not every misinterpretation lends itself to be equally well-presented in both TRUE and FALSE versions and I just want your opinion on the *best* presentation of each (regardless of the resulting distribution of T/F).

To clarify, I am using "TRUE" and "FALSE" for now as placeholders. When we choose the wording we prefer, this will be replaced with VALID/INVALID, CORRECT/INCORRECT, AGREE/DISAGREE, etc. Don't let the use of TRUE/FALSE throw you for the item stems that contain the word "TRUE".

What is your name?



FALSE: The p-value is the probability that the null hypothesis is true.

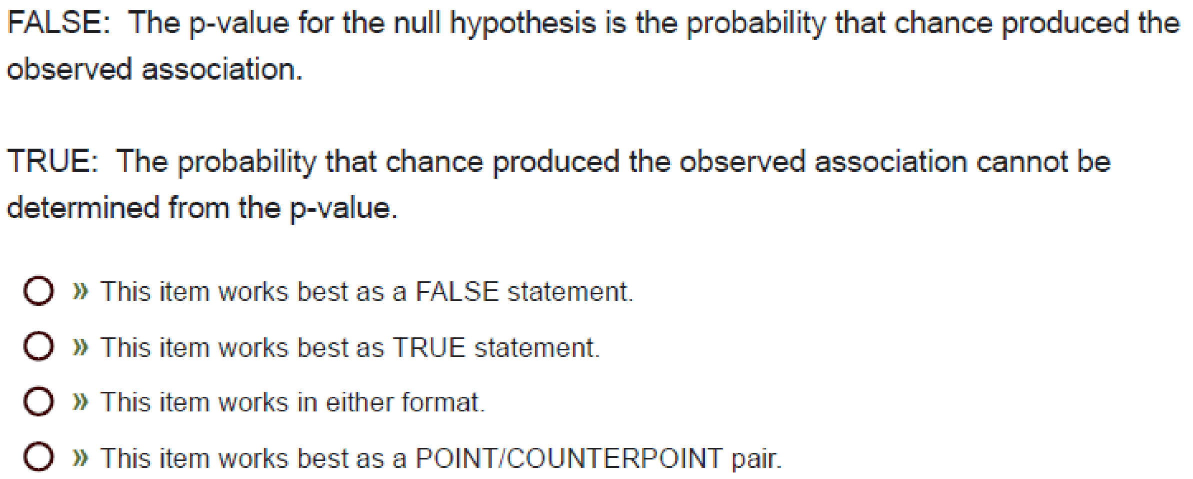
TRUE: The p-value is not a measure of the probability of the truth of the null hypothesis.

○ This item works best as a FALSE statement.

○ This item works best as TRUE statement.

○ This item works in either format.

○ This item works best as a POINT/COUNTERPOINT pair.



FALSE: The p-value for the null hypothesis is the probability that chance produced the observed association.

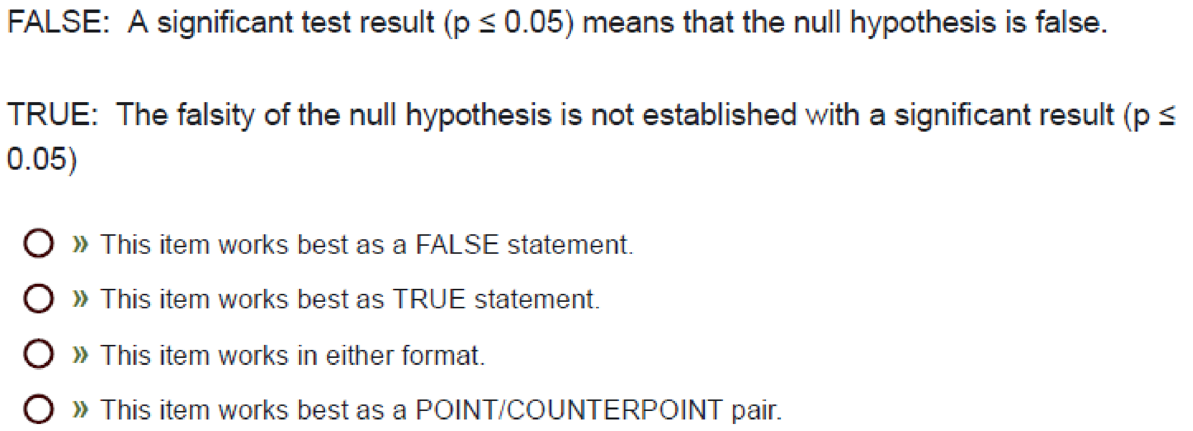
TRUE: The probability that chance produced the observed association cannot be determined from the p-value.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.



FALSE: A significant test result (p ≤ 0.05) means that the null hypothesis is false.

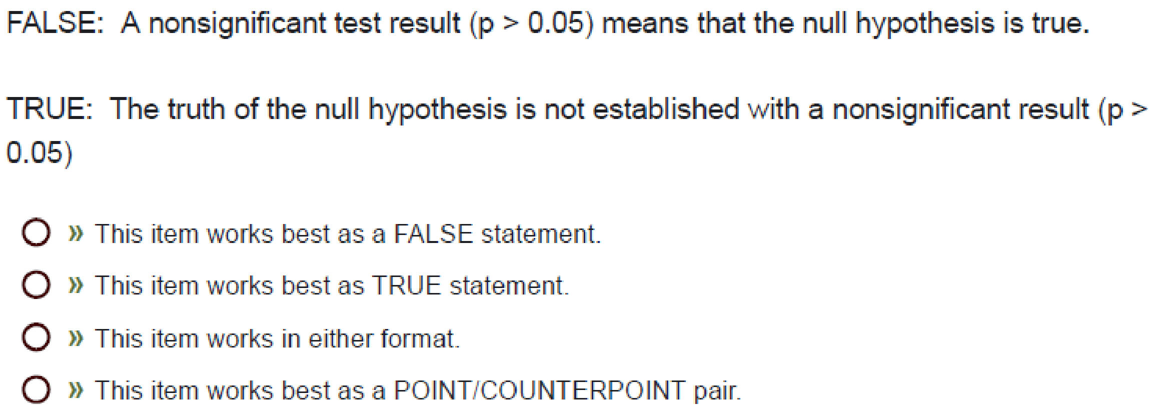
TRUE: The falsity of the null hypothesis is not established with a significant result (p ≤ 0.05)

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.



FALSE: A nonsignificant test result (p > 0.05) means that the null hypothesis is true.

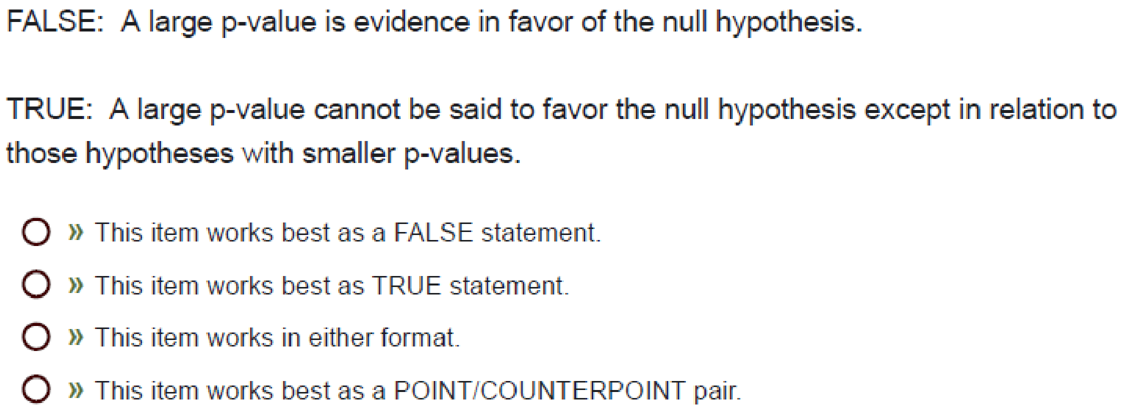
TRUE: The truth of the null hypothesis is not established with a nonsignificant result (p > 0.05)

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.



FALSE: A large p-value is evidence in favor of the null hypothesis.

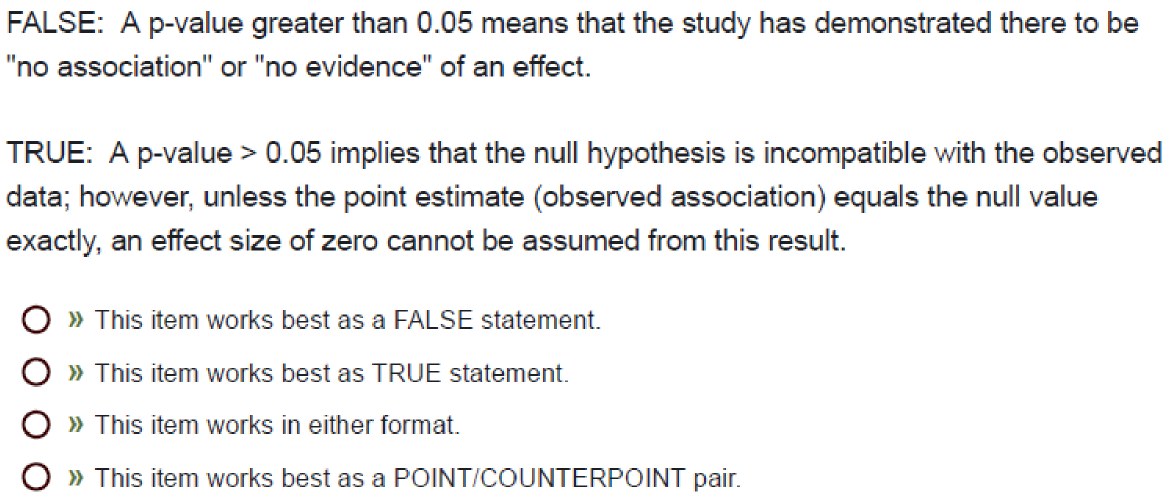
TRUE: A large p-value cannot be said to favor the null hypothesis except in relation to those hypotheses with smaller p-values.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.



FALSE: A p-value greater than 0.05 means that the study has demonstrated there to be "no association" or "no evidence" of an effect.

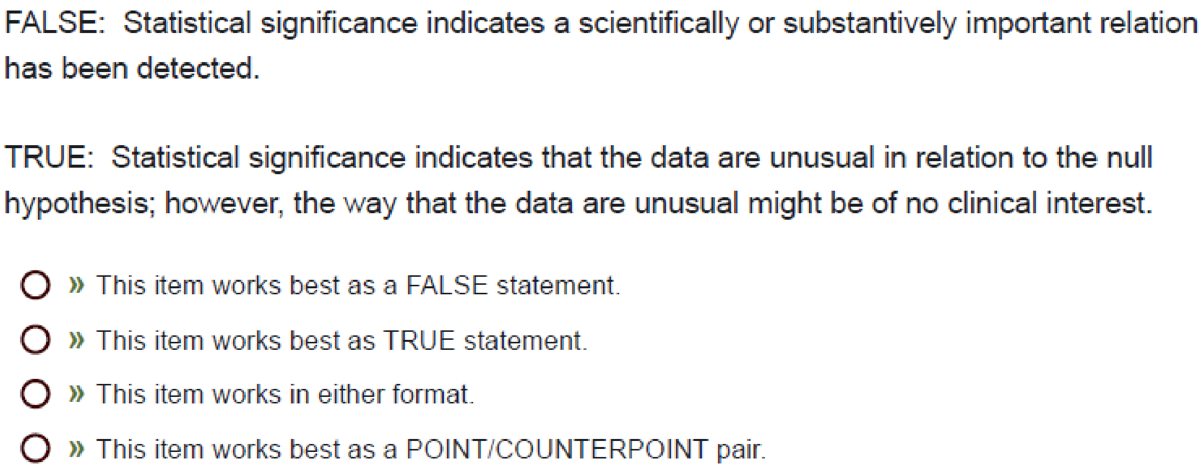
TRUE: A p-value > 0.05 implies that the null hypothesis is incompatible with the observed data; however, unless the point estimate (observed association) equals the null value exactly, an effect size of zero cannot be assumed from this result.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.



FALSE: Statistical significance indicates a scientifically or substantively important relation has been detected.

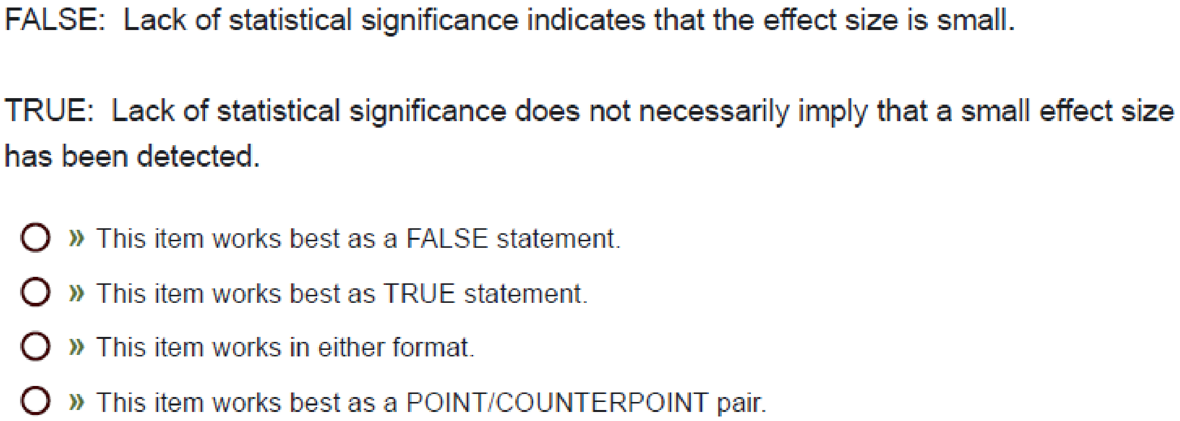
TRUE: Statistical significance indicates that the data are unusual in relation to the null hypothesis; however, the way that the data are unusual might be of no clinical interest.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.



FALSE: Lack of statistical significance indicates that the effect size is small.

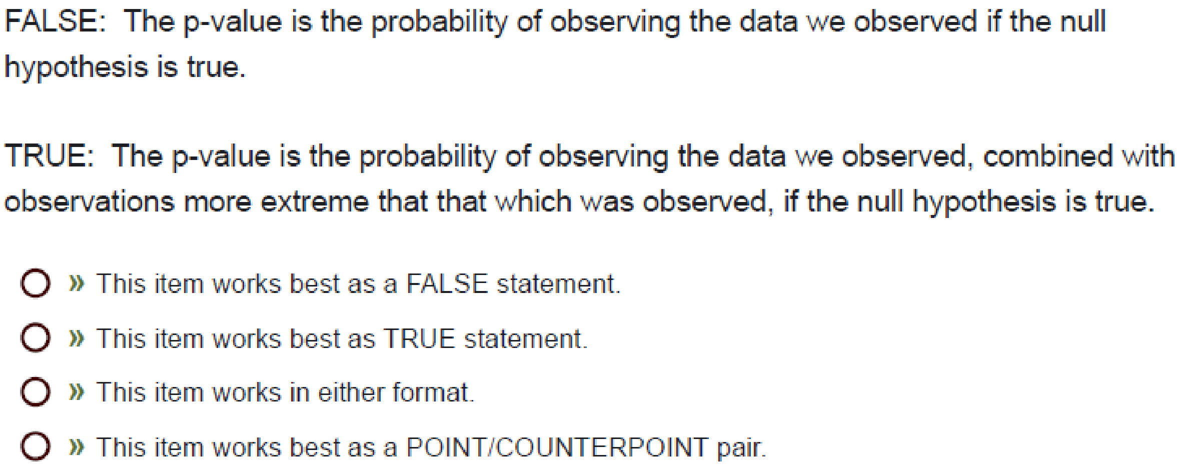
TRUE: Lack of statistical significance does not necessarily imply that a small effect size has been detected.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.



FALSE: The p-value is the probability of observing the data we observed if the null hypothesis is true.

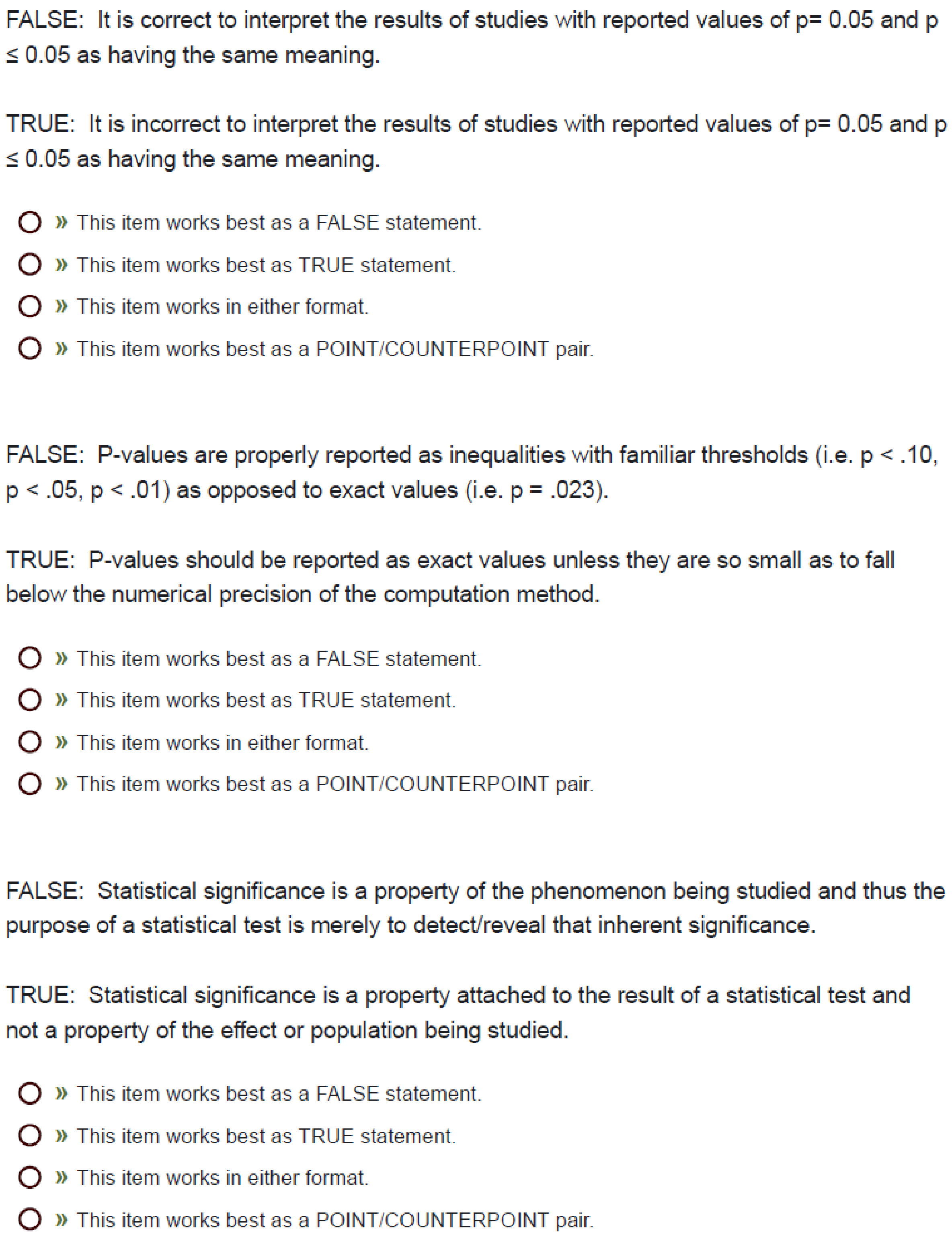
TRUE: The p-value is the probability of observing the data we observed, combined with observations more extreme that that which was observed, if the null hypothesis is true.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.



FALSE: It is correct to interpret the results of studies with reported values of p= 0.05 and p ≤ 0.05 as having the same meaning.

TRUE: It is incorrect to interpret the results of studies with reported values of p= 0.05 and p ≤ 0.05 as having the same meaning.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.

FALSE: P-values are properly reported as inequalities with familiar thresholds (i.e. p < .10, p < .05, p < .01) as opposed to exact values (i.e. p = .023).

TRUE: P-values should be reported as exact values unless they are so small as to fall below the numerical precision of the computation method.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.

FALSE: Statistical significance is a property of the phenomenon being studied and thus the purpose of a statistical test is merely to detect/reveal that inherent significance.

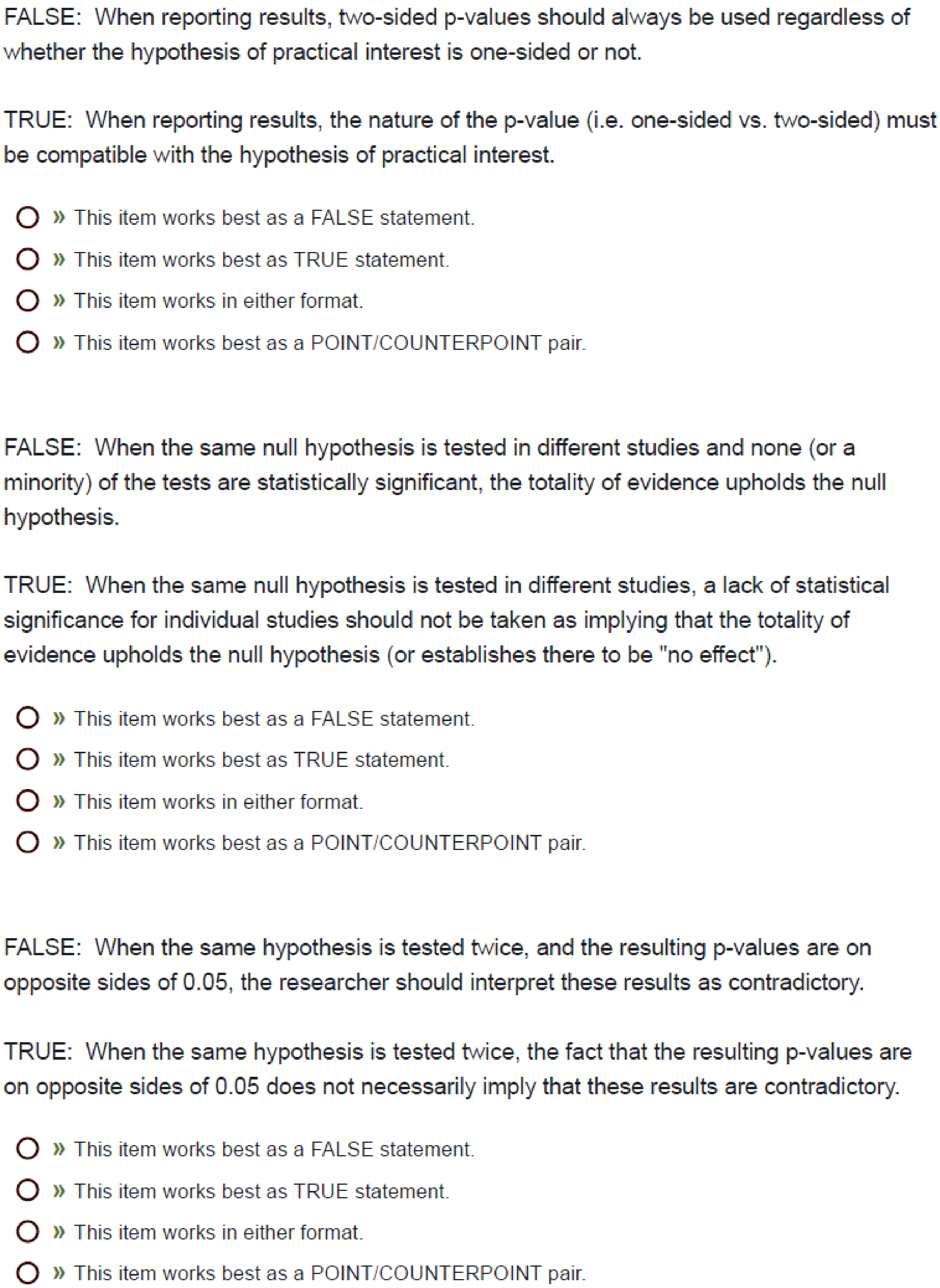
TRUE: Statistical significance is a property attached to the result of a statistical test and not a property of the effect or population being studied.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.



FALSE: When reporting results, two-sided p-values should always be used regardless of whether the hypothesis of practical interest is one-sided or not.

TRUE: When reporting results, the nature of the p-value (i.e. one-sided vs. two-sided) must be compatible with the hypothesis of practical interest.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.

FALSE: When the same null hypothesis is tested in different studies and none (or a minority) of the tests are statistically significant, the totality of evidence upholds the null hypothesis.

TRUE: When the same null hypothesis is tested in different studies, a lack of statistical significance for individual studies should not be taken as implying that the totality of evidence upholds the null hypothesis (or establishes there to be "no effect").

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.

FALSE: When the same hypothesis is tested twice, and the resulting p-values are on opposite sides of 0.05, the researcher should interpret these results as contradictory.

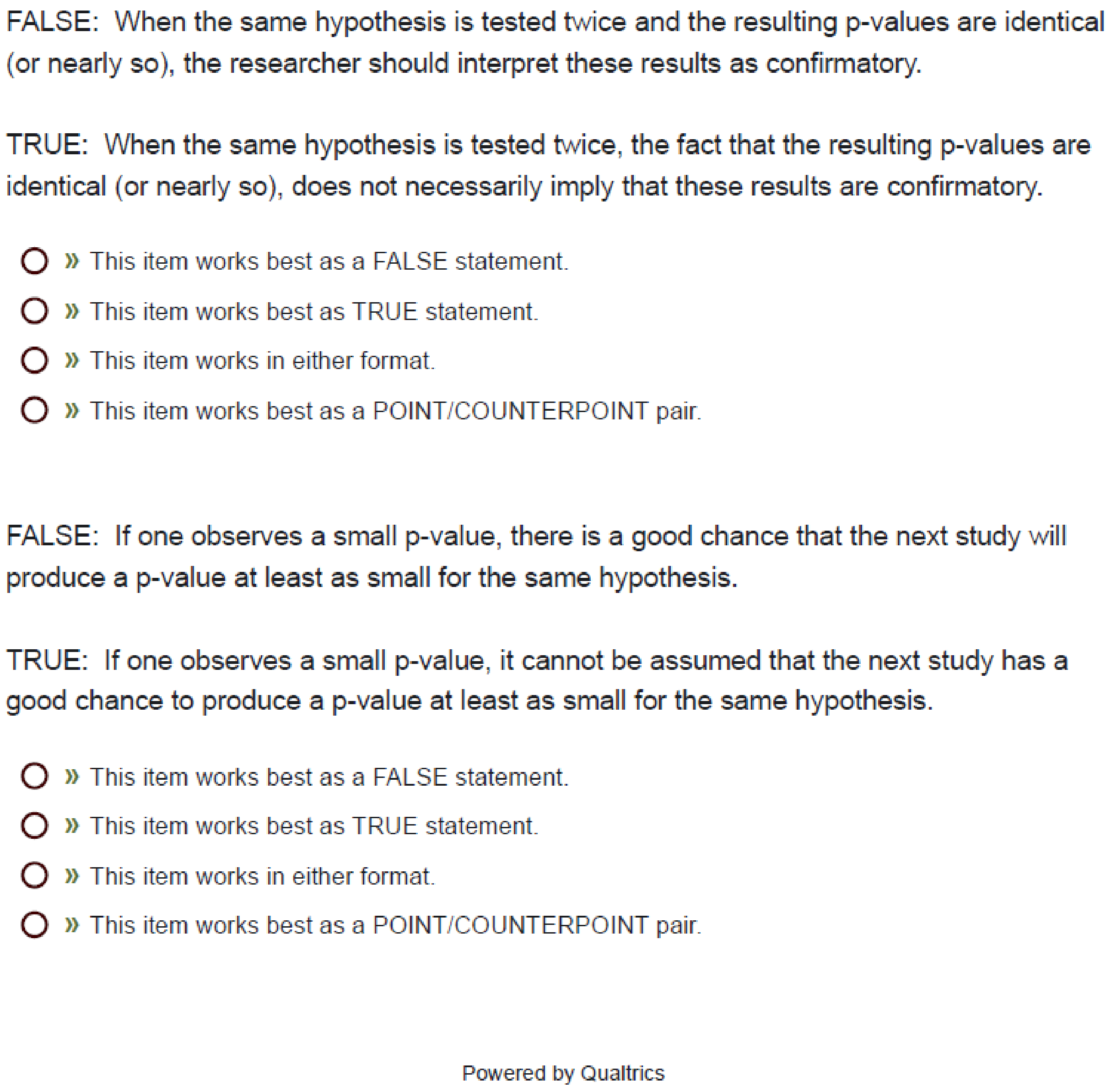
TRUE: When the same hypothesis is tested twice, the fact that the resulting p-values are on opposite sides of 0.05 does not necessarily imply that these results are contradictory.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.



FALSE: When the same hypothesis is tested twice and the resulting p-values are identical (or nearly so), the researcher should interpret these results as confirmatory.

TRUE: When the same hypothesis is tested twice, the fact that the resulting p-values are identical (or nearly so), does not necessarily imply that these results are confirmatory.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.

FALSE: If one observes a small p-value, there is a good chance that the next study will produce a p-value at least as small for the same hypothesis.

TRUE: If one observes a small p-value, it cannot be assumed that the next study has a good chance to produce a p-value at least as small for the same hypothesis.

○ » This item works best as a FALSE statement.

○ » This item works best as TRUE statement.

○ » This item works in either format.

○ » This item works best as a POINT/COUNTERPOINT pair.

Powered by Qualtrics

Appendix H – Interrater Discrepancy Reconciliation

Original (False) Statement	Modified (False) Statement	True Option	Anne	Gary	Decision
1. The P value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave P = 0.01, the null hypothesis has only a 1% chance of being true; if instead it gave P = .40, the null hypothesis has a 40% chance of being true.	The p-value is the probability that the null hypothesis is true.	The p-value is not a measure of the probability of the truth of the null hypothesis.	FALSE	FALSE	FALSE
2. The P value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the P value for the null hypothesis is 0.08, there is an 8% probability that chance alone produced the association.	The p-value for the null hypothesis is the probability that chance produced the observed association.	The probability that chance produced the observed association cannot be determined from the p-value.	FALSE	EITHER	FALSE
3. A significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected.	A significant test result ($p \leq 0.05$) means that the null hypothesis is false.	The falsity of the null hypothesis is not established with a significant result ($p \leq 0.05$)	FALSE	TRUE	PCP
4. A nonsignificant test result ($P > 0.05$) means that the test hypothesis is true or should be accepted.	A nonsignificant test result ($p > 0.05$) means that the null hypothesis is true.	The truth of the null hypothesis is not established with a nonsignificant result ($p > 0.05$)	FALSE	FALSE	FALSE
5. A large P value is evidence in favor of the test hypothesis.	A large p-value is evidence in favor of the null hypothesis.	A large p-value cannot be said to favor the null hypothesis except in relation to those hypotheses with smaller p-values.	FALSE	FALSE	FALSE

Original (False) Statement	**Modified (False) Statement**	**True Option**	**Anne**	**Gary**	**Decision**
6. A null hypothesis P value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.	A p-value greater than 0.05 means that the study has demonstrated there to be "no association" or "no evidence" of an effect.	A p-value > 0.05 implies that the null hypothesis is incompatible with the observed data; however, unless the point estimate (observed association) equals the null value exactly, an effect size of zero cannot be assumed from this result.	FALSE	FALSE	FALSE
7. Statistical significance indicates a scientifically or substantively important relation has been detected.	Statistical significance indicates a scientifically or substantively important relation has been detected.	Statistical significance indicates that the data are unusual in relation to the null hypothesis; however, the way that the data are unusual might be of no clinical interest.	FALSE	TRUE	PCP
8. Lack of statistical significance indicates that the effect size is small.	Lack of statistical significance indicates that the effect size is small.	Lack of statistical significance does not necessarily imply that a small effect size has been detected.	FALSE	TRUE	PCP
9. The P value is the chance of our data occurring if the test hypothesis is true; for example, P = 0.05 means that the observed association would occur only 5% of the time under the test hypothesis.	The p-value is the probability of observing the data we observed if the null hypothesis is true.	The p-value is the probability of observing the data we observed, combined with observations more extreme than that which was observed, if the null hypothesis is true.	EITHER	P/CP	PCP
10. If you reject the test hypothesis because P ≤ 0.05, the chance you are in error (the chance your "significant finding" is a false positive) is 5%.	If you reject the null hypothesis because p ≤ 0.05, the chance you have committed a Type I error (the chance your "significant finding" is a false positive) is 5%.	At this time, none of the TRUE options were considered viable and this item was omitted from the survey.	N/A	N/A	FALSE
11. P = 0.05 and P ≤ 0.05 mean the same thing.	It is correct to interpret the results of studies with reported values of p= 0.05 and p ≤ 0.05 as having the same meaning.	It is incorrect to interpret the results of studies with reported values of p= 0.05 and p ≤ 0.05 as having the same meaning.	EITHER	EITHER	TRUE

Original (False) Statement	Modified (False) Statement	True Option	Anne	Gary	Decision
12. P values are properly reported as inequalities (e.g., report "P < 0.02" when P = 0.015 or report "P > 0.05" when P = 0.06 or P = 0.70)	P-values are properly reported as inequalities with familiar thresholds (i.e. p < .10, p < .05, p < .01) as opposed to exact values (i.e. p = .023).	P-values should be reported as exact values unless they are so small as to fall below the numerical precision of the computation method.	TRUE	TRUE	TRUE
13. Statistical significance is a property of the phenomenon being studied, and thus statistical tests detect significance.	Statistical significance is a property of the phenomenon being studied and thus the purpose of a statistical test is merely to detect/reveal that inherent significance.	Statistical significance is a property attached to the result of a statistical test and not a property of the effect or population being studied.	EITHER	TRUE	TRUE
14. One should always use two-sided P values.	When reporting results, two-sided p-values should always be used regardless of whether the hypothesis of practical interest is one-sided or not.	When reporting results, the nature of the p-value (i.e. one-sided vs. two-sided) must be compatible with the hypothesis of practical interest.	EITHER	TRUE	TRUE
15. When the same hypothesis is tested in different studies and none (or a minority) of the tests are statistically significant (all p > 0.05), the overall evidence supports the hypothesis.	When the same null hypothesis is tested in different studies and none (or a minority) of the tests are statistically significant, the totality of evidence upholds the null hypothesis.	When the same null hypothesis is tested in different studies, a lack of statistical significance for individual studies should not be taken as implying that the totality of evidence upholds the null hypothesis (or establishes there to be "no effect").	EITHER	P/CP	PCP
16. When the same hypothesis is tested in two different populations, and the resulting P values are on opposite sides of 0.05, the results are conflicting.	When the same hypothesis is tested twice, and the resulting p-values are on opposite sides of 0.05, the researcher should interpret these results as contradictory.	When the same hypothesis is tested twice, the fact that the resulting p-values are on opposite sides of 0.05 does not necessarily imply that these results are contradictory.	FALSE	TRUE	PCP
17. When the same hypothesis is tested in two different populations and the same P values are obtained, the results are in agreement.	When the same hypothesis is tested twice and the resulting p-values are identical (or nearly so), the researcher should interpret these results as confirmatory.	When the same hypothesis is tested twice, the fact that the resulting p-values are identical (or nearly so), does not necessarily imply that these results are confirmatory.	FALSE	EITHER	FALSE

Original (False) Statement	Modified (False) Statement	True Option	Anne	Gary	Decision
18. If one observes a small P value, there is a good chance that the next study will produce a P value at least as small for the same hypothesis.	If one observes a small p-value, there is a good chance that the next study will produce a p-value at least as small for the same hypothesis.	If one observes a small p-value, it cannot be assumed that the next study has a good chance to produce a p-value at least as small for the same hypothesis.	FALSE	TRUE	PCP

Appendix I – Initial Vignette Pool

Directions: Read each research vignette and the conclusions. For each statement, decide if the interpretation is a statistical distortion of the scientific results or if the researcher has made a statistically defensible assertion.

Scenario 1A (test for independence): An insurance company wants to know if the color of an automobile has a relationship with incidence of moving violations. Using data obtained from police reports nationwide, a sample of 443 cars was collected in which the color of the car {white, black, red, silver, other} and the number of moving violations {0-1, 2-3, 3+} was recorded. The appropriate test procedure yielded a p-value of $p = .0519$.

H0: car color and incidence of moving violations are independent
Ha: car color and incidence of moving violations are not independent

- #1: There is a 5.19% chance that car color and incidence of moving violations are independent (i.e. the null is true)
- #2: The probability that chance produced the observed association between car color and incidence of moving violations is 5.19%.
- #4: Since $p > 0.05$, this means that the null is true (i.e. car color and incidence of moving violations are independent)
- #6: Since $p > 0.05$, this means that the study has demonstrated there to be no association between car color and incidence of moving violations
- #9: There was a 5.19% chance of observing the association between car color and incidence of moving violations that we observed if these variables are truly independent (i.e. the null is true)

Scenario 1B (test for independence): A television broadcasting company wants to maximize profit by airing the sports contests of the greatest local interest and is trying to determine whether differentiated programming by geographic region is warranted. In a national sample of 336 adults, participants were asked to indicate their region {Northeast, Southeast, Midwest, West} and their favorite sport {football, basketball, baseball, soccer}. The appropriate test procedure yielded a p-value of $p = .0481$.

H0: geographic region and sports preference are independent
Ha: geographic region and sports preference are not independent

- #1: There is a 4.81% chance that car color and incidence of moving violations are independent (i.e. the null is true)
- #2: The probability that chance produced the observed association between car color and incidence of moving violations is 4.81%.
- #10: The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error (i.e. rejecting the claim of independence when it should not have been rejected) in this instance is 5%.
- #3: Since $p < 0.05$, this means that the null is false (i.e. geographic region and sports preference are not independent)
- #9: There was a 4.81% chance of observing the association between car color and incidence of moving violations that we observed if these variables are truly independent (i.e. the null is true)

Scenario 2A (1-sample test, 1-sided hypothesis): Faculty are irritated by students' cell phones and claim that they ring during class an average of 15 times per semester. A reporter for the school newspaper claims that students are increasingly more courteous with their phones and that the disruptions to class have been reduced. A random sample of 12 professors found an average of 14.8 calls with a standard deviation of 1.2 calls. The (one-tailed) p-value for the appropriate test procedure was p = .288.

H0: the average number of cell phone disruptions is 15 times per semester
Ha: the average number of cell phone disruptions is less than 15 times per semester

- #1: There is a 28.8% chance that the number of cell phone disruptions is at least 15 times per semester.
- #4: A nonsignificant result here (p = .288 > 0.05) means that the faculty claim (i.e. the null) is true and that the average number of cell phone disruptions is at least 15 times per semester.
- #5: Because this test produced a large p-value, this should be interpreted as evidence in favor of the professors' claim (i.e. incidence of cell phone disruptions is at least 15 times per semester).
- #12: This p-value should properly be reported as an inequality, i.e. p > .05 or p > .10 (as opposed to p = .288).
- #14: This one-sided p-value (p = .288) should be converted to a two-sided p-value when reporting the results of the study.
- #9: The probability of observing the data we did (i.e. an average of 14.8 calls/semester), if the null were true (i.e. average number of cell phone disruptions is actually ≥ 15 per semester), is 28.8%.

Scenario 2B (1-sample, 1-sided hypothesis): Despite the cable company's claims of excellent customer service, most users report that the technicians do not report within the advertised 30-minute window. A random sample of 9 customers found an average wait of 33.2 minutes with a standard deviation of 3.4 minutes. The (one-tailed) p-value for the appropriate test procedure was p = .011.

H0: the average wait time is 30 minutes or less
Ha: the average wait time is greater than 30 minutes

- #1: There is a 1.1% chance that the average customer wait time is 30 minutes or less.
- #9- The probability of observing the data we observed (i.e. average wait of 33.2 minutes), if the null were true (i.e. wait time is ≤ 30 minutes), is 1.1%.
- #10: The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error (i.e. rejecting the company's claim of wait times ≤ 30 minutes when it should not have been rejected) in this instance is 5%.
- #11: If one researcher had reported the results of this study as $p < .05$ and another reported p = .011, it would be correct to interpret these representations as having the same meaning.
- #12: This p-value should properly be reported as $.01 < p < .05$ (as opposed to p = .011).
- #14: This one-sided p-value (p = .011) should be converted to a two-sided p-value when reporting the results of the study.
- #3: A significant test result ($p < .05$) here means that the company's claim of an average customer wait time of 30 minutes or less is false.
- #18: Since a small p-value was observed in this study (p = .011), there is a good chance that the next study testing the same hypothesis (i.e. customer wait times ≤ 30 minutes) will produce a p-value at least as small.
- #16: In a second random sample of 5 customers (average wait time of 33.1 minutes, standard deviation = 3.4 minutes), the p-value for the appropriate test procedure was 0.056. Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the researcher should interpret these results as contradictory/conflicting.
- #17: In a second random sample of 70 customers (average wait time of 30.6 minutes, standard deviation = 2.25 minutes), the p-value for the appropriate test procedure was 0.014. Since the same hypothesis was tested twice and the resulting p-values are nearly identical, the researcher should interpret these results as confirmatory (in agreement?).

Scenario 3A (1-sample, 2-sided hypothesis): A manufacturer must test that his bolts are 2.00 cm long when they come off the assembly line Bolts that are too long or too short cannot be used and the machines must be recalibrated in the event that unsatisfactory bolts are being produced. A random sample of 100 bolts yields a sample mean of 1.90 cm with a standard deviation of 0.5cm. The p-value for the appropriate test procedure was $p = .048$.

H0: the average bolt length is 2.0 cm
Ha: the average bolt length is not equal to 2.0 cm

- #1: The probability that the average bolt length is 2.0 cm (i.e. the null is true) is 4.8%.
- #2: The probability that chance produced the observed association (i.e. a sample mean of 1.9 cm) is 4.8%.
- #10: The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error (i.e. rejecting the claim that the bolts average 2.0 cm when in fact this is true and should not be rejected) in this instance is 5%.
- #17: In a second random sample of 100 bolts (mean = 2.1 cm, standard deviation = 0.5 cm), the p-value for the appropriate test procedure was $p = .048$. Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as confirmatory.
- #17: In a second random sample of 100 bolts (mean = 2.1 cm, standard deviation = 0.5 cm), the p-value for the appropriate test procedure was $p = .048$. Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as being in agreement.
- #17: In a second random sample of 400 bolts (mean = 1.95 cm, standard deviation = 0.5 cm), the p-value for the appropriate test procedure was $p = .046$. Since the same hypothesis was tested twice and the resulting p-values are nearly identical, the manufacturer should interpret these results as being in agreement.
- #12: This p-value should properly be reported as $p < .05$ (as opposed to $p = .011$).
- #3: A significant test result ($p = .048 < .05$) means that the average bolt length is not equal to 2.0 cm (i.e. that the null is false).
- #9: The probability of observing the data we did (i.e. an average bolt length of 1.9 cm), if the null were true (i.e. average bolt length is 2.0 cm), is 4.8%.
- #11: If one researcher had reported the results of this study as $p = .048$ and another reported $p < .05$, it would be correct to interpret these representations as having the same meaning.
- #16: In a second random sample of 85 bolts (mean = 1.90 cm, standard deviation = 0.5 cm), the p-value for the appropriate test procedure was $p = .068$. Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the manufacturer should interpret these results as conflicting.
- #18: Since a small p-value was observed in this study ($p = .048$), there is a good chance that the next study testing the same hypothesis (i.e. average bolt length = 2.0 cm) will produce a p-value at least as small.

Scenario 3B (1-sample, 2-sided hypothesis): A manufacturer is responsible for making barrels used to store crude oil, which are designed to hold exactly 55 gallons of oil. As part of a routine quality check, a factory manager randomly tests 27 barrels off the assembly line and finds the mean capacity is 54.7 gallons with a standard deviation of 0.8 gallons. The p-value for the appropriate test procedure was .062.

H0: the average barrel capacity is 55 gallons
Ha: the average barrel capacity is not equal to 55 gallons

- #1: The probability that the average barrel capacity is 55 gallons (i.e. the null is true) is 6.2%.
- #2: The probability that chance produced the observed association (i.e. a sample mean of 54.7 gal) is 6.2%.
- #4: A nonsignificant test result (p = .062 > 0.05) means that the average barrel capacity is 55 gallons (i.e. the null is true).
- #17: In a second random sample of 27 barrels (mean = 55.3 gallons, standard deviation = 0.8 gal), the p-value for the appropriate test procedure was p = .062. Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as confirmatory.
- #17: In a second random sample of 27 barrels (mean = 55.3 gallons, standard deviation = 0.8 gal), the p-value for the appropriate test procedure was p = .062. Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as being in agreement.
- #17: In a second random sample of 100 barrels (mean = 54.85 gal, standard deviation = 0.8 gal), the p-value for the appropriate test procedure was p = .0637. Since the same hypothesis was tested twice and the resulting p-values are nearly identical, the manufacturer should interpret these results as being in agreement.
- #12: This p-value should properly be reported as .05 < p < .10 or p < .10 (as opposed to p = .062).
- #9: The probability of observing the data we did (i.e. an average barrel capacity of 54.7 gallons), if the null were true (i.e. average barrel capacity is 55 gallons), is 6.2%.
- #16: In a second random sample of 35 barrels (mean = 54.7 gal, standard deviation = 0.8 gal), the p-value for the appropriate test procedure was p = .033. Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the factory manager should interpret these results as conflicting.
- #15: In a second random sample of 25 barrels (mean = 54.7 gal, standard deviation = 0.8 gal), the p-value for the appropriate test procedure was p = .073. Since the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager should conclude that the totality of evidence upholds the null hypothesis (i.e. the claim that the average barrel capacity is 55 gallons).

Scenario 4A (2-sample independent, 1-sided hypothesis): A parent interest group is looking at whether birth order affects scores on the ACT test. It was suggested that first-born children earn lower scores than second-born children. A survey of 100 first-born and 175 second-born children resulted in mean scores of 20.9 (standard deviation = 1.8) and 21.1 (standard deviation 2.3), respectively. The (one-tailed) p-value for the appropriate test procedure was p = .212.

H0: first-born children score as well (or better) on the ACT test than do second-born children
Ha: first-born children earn lower scores on the ACT test than do second-born children

- #1: The probability that first-born children score as well on the ACT test as do second-born children is 21.2%.
- #2: The probability that chance produced the observed association (i.e. lower scores in first-born children in our sample) is 21.2%.
- #4: Since our test result was non-significant ($p > 0.05$), this means that first-born children do not earn lower scores on the ACT as compared with second-born children (i.e. the null is true).
- #5: Because this test produced a large p-value, this should be interpreted as evidence in favor of the claim that first-born children score as well (or better) on the ACT test than do second-born children.
- #10: The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error (i.e. rejecting the claim that first-born children score as well (or better) on the ACT than do second-born children when in fact this is true and should not be rejected) in this instance is 5%.
- #12: The p-value should be properly reported as $p > .05$ or $p > .10$ (as opposed to $p = .212$).
- #14: This one-sided p-value ($p = .212$) should be converted to a two-sided value when reporting the results of the study.
- #9: The probability of observing the data we did (i.e. a 0.2 point difference in mean ACT score in favor of second-born children), if the null hypothesis were true (i.e. first-born children score as well or better than do second-born children), is 21.2%.

Scenario 4B (2-sample independent, 1-sided hypothesis): A weight-loss company wants to make sure that its clients lose more weight, on average, than they would without the company's help. An independent researcher collects data on 25 clients to compare to a control group of 50 people not using the company's program. Mean weight loss was 12 pounds (std. dev. of 5 lb) for the treatment group, 10 pounds (std.dev. of 6 lb) for the control group. The (one-tailed) p-value for the appropriate test procedure was p = .07.

H0: clients who use the program lose no more weight on average than those who don't
Ha: clients who use the program lose more weight on average than those who don't

- #1: The probability that clients using the program lose no more weight on average than those who don't (i.e. the null is true) is 7%.
- #2: The probability that chance produced the observed association (i.e. higher weight loss in the treatment group) is 7%.
- #4: Since our test result was non-significant (p > 0.05), this means that clients who use the program lose no more weight on average than those who don't (i.e. the null is true).
- #6: Since our p-value was greater than 0.05, this means that the study has demonstrated the company's program to have no effect on weight loss.
- #12: The p-value should be properly reported as .05 < p < .10 (as opposed to p = .07).
- #14: This one-sided p-value (p = .07) should be converted to a two-sided value when reporting the results of the study.
- #8: Lack of statistical significance (i.e. p > 0.05) indicates that the effect size (i.e. the program's influence on weight loss) is small.
- #9: The probability of observing the data we did (i.e. a 2-lb difference in weight loss between the treatment and control groups), if the null hypothesis were true (i.e. clients on the program lose no more weight than those not on it), is 7%.
- #15: In a second random sample of 15 clients (mean = 12.5 lb, standard deviation = 5 lb), the p-value for the appropriate test procedure was p = .059. Since the same hypothesis was tested twice and neither of the tests was statistically significant, the weight loss company should conclude that the totality of evidence upholds the null hypothesis (i.e. the claim that clients who use the program lose no more weight on average than those who don't).

Scenario 5A (2-sample dependent, 1-sided hypothesis): A local school district is looking at adopting a new textbook that, according to the publishers, will increase standardized test scores of second graders by more than 10 points, on average. Never willing to believe a publisher's claim without evidence to support it, the school board decides to test the claim. Two classrooms are selected for the study, one of which is assigned the new textbook and the other used the traditional text. 8 children from each class were paired based on demographics and ability level resulting in a mean difference per pair of 11.5 points with standard deviation of .76 points. The (one-tailed) p-value for the appropriate test procedure was p = .00042.

> H0: Students using the new textbook will show standardized test score gains of ≤ 10 points.
> Ha: Students using the new textbook will show standardized test score gains of at least 10 points.

- #1: The probability that students using the new textbook will show standardized test score gains of ≤ 10 points. (i.e. the null is true) is .042%.
- #2: The probability that chance produced the observed association (i.e. higher scores for the students using the new text) is .042%.
- #10: The null hypothesis was rejected since p < 0.05; therefore, the chance the researcher committed a Type I error (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) in this instance is 5%.
- #12: The p-value should be properly reported as p < .01 or p < .001 (as opposed to p = .00042).
- #13: Statistical significance is a property of the effect textbook has on the test scores and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.
- #14: This one-sided p-value (p = .00042) should be converted to a two-sided value when reporting the results of the study.
- #3: A significant test result here (p = .00042 < .05) means that the null is false (i.e. the true difference in test scores between the groups is not ≤ 10 points).
- #7: Statistical significance (p = .00042 < .05) indicates that a scientifically or substantively important relation between textbook and test scores has been detected.
- #9: The probability of observing the data we did (i.e. 11.5 point score difference), if the null hypothesis were true (i.e. the difference is ≤ 10 points), is .042%.
- #11: If one researcher had reported the results of this study as p = .00042 and another reported p < .05, it would be correct to interpret these representations as having the same meaning.
- #18: Since a small p-value was observed in this study (p = .00042), there is a good chance that the next study testing the same hypothesis (i.e. increase in test scores ≤ 10 points) will produce a p-value at least as small.

Scenario 5B (2-sample dependent, 1-sided hypothesis): An SAT prep course claims to increase scores by more than 60 points, on average. To test this claim, 9 students who have previously taken the SAT are randomly chosen to take the prep course. Their SAT scores before and after completing the prep course are compared and a mean difference of 81 points is found (standard deviation of 87 points). The p-value for the appropriate test procedure was p = .2436.

H0: Students who take the prep course improve their SAT scores by 60 points or less.
Ha: Students who take the prep course improve their SAT scores by more than 60 points.

- #1: The probability that students taking the prep course will improve their SAT score by ≤ 60 points. (i.e. the null is true) is 24.36%.
- #2: The probability that chance produced the observed association (i.e. improved scores by 81 points for those students in the prep course) is 24.36%.
- #4: A nonsignificant result here (i.e. p = .2436 > .05) means that the null is true (i.e. students who take the prep course improve their SAT score by 60 points or less).
- #5: Because this test produced a large p-value (p = .2436), this should be interpreted as evidence in favor of the claim that students who take the prep course improve their SAT score by 60 points or less
- #6: Since our p-value was greater than 0.05, this means that the study has demonstrated there to be no association between the prep course and score improvements greater than 60 points (in other words, the study has demonstrated, by virtue of the size of the p-value, that there is no evidence of a 60-point improvement effect attributable to the prep course).
- #12: The p-value should be properly reported as p > .05 or p > .10 (as opposed to p = .2436).
- #13: Statistical significance is a property of the effect the prep course has on SAT scores and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.
- #14: This one-sided p-value (p = .2436) should be converted to a two-sided value when reporting the results of the study.
- #8: Lack of statistical significance (p = .2436 > .05) indicates that the effect size (of the prep course on SAT scores) is small.
- #9: The probability of observing the data we did (i.e. 81 point improvement on SAT scores after prep course), if the null hypothesis were true (i.e. true improvement is ≤ 60 points), is 24.36%.

Scenario 6A (2-sample independent, 1-sided hypothesis): There is an old wives' tale that women who eat chocolate during pregnancy are more likely to have happy babies. To test this claim, 100 randomly selected pregnant women are recruited and half agree to eat chocolate at least once a day while the other half agree to forego chocolate for the duration of their pregnancies. A year later, the women complete a survey regarding the overall happiness of their babies in which the number of happy babies reported is 24 and 22 for the chocolate/not groups, respectively. The (one-sided) p-value for the appropriate test procedure was p = .344.

H0: women who eat chocolate are not more likely to have happy babies
Ha: women who eat chocolate are more likely to have happy babies

- #1: The probability that women who eat chocolate are not more likely to have happy babies (i.e. the null is true) is 34.4%.
- #2: The probability that chance produced the observed association between baby happiness and chocolate consumption (i.e. 4% more happy babies in chocolate group) is 34.4%.
- #4: A nonsignificant result here (p = .344 > 0.05) means that the null is true (i.e. women who eat chocolate are NOT more likely to have happy babies).
- #5: Because this test produced a large p-value, this should be interpreted as evidence in favor of the position that women who eat chocolate are not more likely to have happy babies.
- #6: Since the p-value was greater than 0.05, this means that the study has demonstrated there to be no association between chocolate consumption and the temperament of babies.
- #12: The p-value should be properly reported as p > .10 (as opposed to p = .344).
- #14: This one-sided p-value (p = .344) should be converted to a two-sided value when reporting the results of the study.
- #8: Lack of statistical significance (p = .344 > 0.05) in this study indicates that the effect size (i.e. effect of chocolate consumption on the temperament of babies) is small.
- #9: The probability of observing the data we observed (i.e. 4% more happy babies in the chocolate group), if the null hypothesis were true (i.e. women who eat chocolate are not more likely to have happy babies), is 34.4%.

Scenario 6B (2-sample independent, 1-sided hypothesis): Researchers claim that the birth rate in Bonn, Germany is higher than the national average. A random sample of 900 Bonn residents produced 19 births, whereas a random sample of 1100 people from all over Germany had 12 births during the same year. The p-value for the appropriate test procedure was .033.

H0: the birth rate in Bonn is not higher than the national average
Ha: the birth rate in Bonn is higher than the national average

- #1: The probability that the birth rate in Bonn is not higher than the national average (i.e. the null is true) is 3.3%.
- #2. The probability that chance produced the observed association (i.e. birth rate of 19/900 in Bonn vs. 12/1100 nationwide) is 3.3%.
- #10: The null hypothesis was rejected since p < 0.05; therefore, the chance the researcher committed a Type I error (i.e. rejecting the claim that the birth rate in Bonn is not higher than the national average when this is true) in this instance is 5%.
- #12: The p-value should properly be reported as .01 < p < .05 or p < .05 (as opposed to p = .033).
- #14: This one-sided p-value (p = .033) should be converted to a two-sided value when reporting the results of the study.
- #7: Statistical significance in this study (p = .033 < .05) indicates that a scientifically or substantively important relation between birth rate and Bonn has been detected.
- #9: The probability of observing the data we observed (i.e. birth rate of 19/900 in Bonn vs. 12/1100 nationwide), if the null were true (i.e. the birth rate in Bonn is not higher than the national average), is 3.3%.
- #11: If one researcher had reported the results of this study as p = .033 and another reported p < .05, it would be correct to interpret these representations as having the same meaning.
- #18: Since a small p-value was observed in this study (p = .033), there is a good chance that the next study testing the same hypothesis (i.e. birth rate in Bonn is not higher than the national average) will produce a p-value at least as small.

Scenario 7A (2-sample independent, 2-sided hypothesis): ????do we need this

Scenario 7B (2-sample independent, 2-sided hypothesis):

Scenario 8A (ANOVA): A casino manager is concerned that one of the blackjack tables is not generating the same revenue, on average, as the other three tables. The amount of revenue received during one particular three-hour period is recorded at each of four tables each night for five consecutive nights resulting in a grand mean of $8667.90 and table means of $8486.60, $8464.20, $9509.40, and $7911.40 for Tables 1, 2, 3, & 4, respectively. The p-value for the appropriate test procedure was p = .0817.

H0: all tables generate the same revenue on average
Ha: there is at least one table for which the average revenue differs from the others

- #1: The probability that all tables generate the same revenue on average (i.e. the null is true) is 8.17%.
- #2: The probability that chance produced the observed association (Table 4 < Table 2 < Table 1 < Table 3), if the null were true (i.e. all tables generate the same revenue on average), is 8.17%.
- #4: A nonsignificant results here (p = .0817 > 0.05) means that the null is true (i.e. all tables generate the same revenue on average).
- #6: A p-value greater than 0.05 here means that the study has demonstrated there to be no association between shift and call volume.
- #8: Lack of statistical significance (i.e. p = .0817 > 0.05) indicates that the effect size (effect of shift on call volume) is small.
- #12: The p-value should be properly reported as p < .10 or .05 < p < .10 (as opposed to p = .0817).
- #9: The probability of observing the data we observed (table means of $8486.60, $8464.20, $9509.40, and $7911.40), if the null were true (i.e. all tables generate the same revenue on average), is 8.17%.
- #15: The following week, the sampling is repeated for the same three-hour period for the same four tables for five consecutive nights yielding table means of $8470.40, $8466, $9498.60, $7916.80, and $8587.95 with a p-value of p = .082. Since the same hypothesis was tested twice and neither of the tests was statistically significant, the casino manager should conclude that the totality of evidence upholds the null hypothesis (i.e. the claim that all tables generate the same revenue on average).

Scenario 8B (ANOVA): A group of paramedics does not believe that the mean number of calls received in one shift is the same for the morning, afternoon, and evening shifts. To test this claim, they record the number of calls received during each shift for seven days, resulting in the following call/shift averages: 2.57, 3.71, and 4.28 for the morning, afternoon, and evening shifts, respectively. The p-value for the appropriate test procedure was p = .039.

H0: all shifts receive the same number of calls on average
Ha: there is at least one shift for which the average number of calls differs from the others

- #1: The probability that all shifts receive the same number of calls on average (i.e. the null is true) is 3.9%.
- #2: The probability that chance produced the observed association (i.e. morning < afternoon < evening) is 3.9%.
- #10: The null hypothesis was rejected since p < 0.05; therefore, the chance the researcher committed a Type I error (i.e. rejecting the claim that all shifts receive the same number of calls on average when in fact they do) in this instance is 5%.
- #12: The p-value should be properly reported as p < .05 or .01 < p < .05 (as opposed to p = .039).
- #13: Statistical significance is a property of the effect time of day has on number of calls received and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.
- #3: A significant results here (p = .039 < 0.05) means that the null is false (i.e. that all shifts do NOT receive the same number of calls on average).
- #7: Statistical significance (p = .039 < 0.05) means that a substantively important relation between shift and call volume has been detected.
- #9: The probability of observing the data we observed (shift means of 2.57, 3.71, and 4.28), if the null was true (i.e. all shifts receive the same number of calls on average), is 3.9%.
- #11: If one researcher had reported the results of this study as p = .039 and another reported p < .05, it would be correct to interpret these representations as having the same meaning.
- #18: Since a small p-value was observed in this study (p = .039), there is a good chance that the next study testing the same hypothesis (i.e. all shifts receive the same number of calls on average) will produce a p-value at least as small.

Appendix J – Vignettes with Edits

Scenario 2A: Professors are irritated by students' cell phones and claim that they ring during class an average of 15 times per semester. A reporter for the school newspaper claims that students are increasingly more courteous with their phones and that the disruptions to class have been reduced. A random sample of 12 professors found an average of 14.8 calls with a standard deviation of 1.2 calls. The (one-tailed) p-value for the appropriate test procedure was p = .288.

Comment [GS1]: Is 15 a population mean or a sample mean? I think you mean the former, but this reads like the latter. How would faculty know this figure? Suggest that 15 comes from a national survey, census, or other authoritative source.

H0: the average number of cell phone disruptions is at least 15 times per semester
Ha: the average number of cell phone disruptions is less than 15 times per semester

- #1:
  - There is a 28.8% chance that the number of cell phone disruptions is at least 15 times per semester (i.e. that the null is true).
  - The probability that the number of cell phone disruptions is at least 15 times per semester (i.e. the null is true) cannot be determined from the p-value.
- #4:
  - A nonsignificant result here (p = .288 > 0.05) means that the professors' claim (i.e. the null) is true and that the average number of cell phone disruptions is at least 15 times per semester.
  - The truth of the faculty's claim (i.e. the null hypothesis) is not established with a non-significant result.
- #5:
  - Because this test produced a large p-value, this should be interpreted as evidence in favor of the professors' claim (i.e. incidence of cell phone disruptions is at least 15 times per semester).
  - A large p-value cannot be said to favor the faculty's claim (i.e. this specific null hypothesis) except in relation to those hypotheses with smaller p-values.

Comment [AD2]: I am confused on what you mean here, "except in relation to those hypotheses with smaller p-values"?

- #14:
  - This one-sided p-value (p = .288) should be converted to a two-sided p-value when reporting the results of the study.
  - This one-sided p-value (p = .288) should be reported as-is in order to maintain compatibility with the hypothesis of practical interest.
- #9:
  - The probability of observing the data we did (i.e. an average of 14.8 calls/semester), if the null were true (i.e. average number of cell phone disruptions is actually ≥ 15 per semester), is 28.8%.
  - The probability of observing data at least as extreme (i.e. an average ≥ 14.8 calls/semester) as that which was observed, if the null were true (i.e. average number of cell phone disruptions is actually ≥ 15 per semester), is 28.8%.

Scenario 2B: Despite the cable company's claims of excellent customer service, most users report that the technicians do not report within the advertised 30-minute window. A random sample of 9 customers found an average wait of 33.2 minutes with a standard deviation of 3.4 minutes. The (one-tailed) p-value for the appropriate test procedure was $p = .011$.

Comment [AD3]:

Comment [AD4]: Maybe put "find" here because you use report twice in the same sentence.

H0: the average wait time is 30 minutes or less
Ha: the average wait time is greater than 30 minutes

- #10:
  - The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error (i.e. rejecting the company's claim of wait times ≤ 30 minutes when it should not have been rejected) in this instance is 5%.
  - The probability of committing a Type I error (i.e. rejecting the company's claim of wait times ≤ 30 minutes when the claim is valid and should not have been rejected) is a long-run frequency equivalent to the significance level of the test (i.e., $\alpha = 0.05$) and indeterminable for any particular hypothesis test.
- #11:
  - If one researcher had reported the results of this study as $p < .05$ and another reported $p = .011$, it would be correct to interpret these representations as having the same meaning.
  - Reported p-values of $p = .011$ and $p < 0.05$ do not convey the same meaning and should not be interpreted as equivalent.
- #18:
  - Since a small p-value was observed in this study ($p = .011$), there is a good chance that the next study testing the same hypothesis (i.e. customer wait times ≤ 30 minutes) will produce a p-value at least as small.
  - Just because a small p-value was observed in this study ($p = .011$), it cannot be assumed that the next study has a good chance to produce a p-value at least as small for the same hypothesis.
- #16:
  - Suppose a second random sample of 5 customers (average wait time of 33.1 minutes, standard deviation = 3.4 minutes) yielded a p-value for the appropriate test procedure of 0.056. Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the researcher should interpret these results as contradictory/conflicting.
  - Suppose a second random sample of 5 customers (average wait time of 33.1 minutes, standard deviation = 3.4 minutes) yielded a p-value for the appropriate test procedure of 0.056. Despite the fact that the resulting p-values are on opposite sides of 0.05, the researcher could interpret these results as confirmatory/corroborating.
- #17:
  - Suppose a second random sample of 70 customers (average wait time of 30.6 minutes, standard deviation = 2.25 minutes) yielded a p-value for the appropriate test procedure of $p = 0.014$. Since the same hypothesis was tested twice and the resulting p-values are nearly identical, the researcher should interpret these results as confirmatory (in agreement?).
  - Suppose a second random sample of 70 customers (average wait time of 30.6 minutes, standard deviation = 2.25 minutes) yielded a p-value for the appropriate test procedure of $p = 0.014$. Despite the fact that the resulting p-values are nearly identical, the researcher should not assume these results to be in agreement.

Comment [AD5]: Do you need the ?

Scenario 3A: A manufacturer must test that his bolts are 2.00 cm long when they come off the assembly line. –Bolts that are too long or too short cannot be used and the machines must be recalibrated in the event that unsatisfactory bolts are being produced. A random sample of 100 bolts yields a sample mean of 1.90 cm with a standard deviation of 0.5cm. The p-value for the appropriate test procedure was p = .048.

Comment [AD6]:

H0: the average bolt length is 2.0 cm
Ha: the average bolt length is not equal to 2.0 cm

- #1:
  - The probability that the average bolt length is 2.0 cm (i.e. the null is true) is 4.8%.
  - The probability that the average bolt length is 2.0 cm (i.e. the null is true) cannot be determined from the p-value.

Comment [GS7]: Probabilities range from 0 to 1. Respondents could disagree because probabilities are not expressed as percents.

- #17:
  - Suppose a second random sample of 100 bolts (mean = 2.1 cm, standard deviation = 0.5 cm) yielded a p-value for the appropriate test procedure of p = .048. Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as confirmatory (in agreement?).
  - Suppose a second random sample of 100 bolts (mean = 2.1 cm, standard deviation = 0.5 cm) yielded a p-value for the appropriate test procedure of p = .048. Despite the fact that the resulting p-values are identical, the researcher should not assume these results to be in agreement.

- #3:
  - A significant test result (p = .048 < .05) means that the average bolt length is not equal to 2.0 cm (i.e. that the null is false).
  - The falsity of the null hypothesis (i.e. the average bolt length is not equal to 2.0 cm) is not established with a significant result (i.e. p = .048 < 0.05).
- #9:
  - The probability of observing the data we did (i.e. an average bolt length of 1.9 cm), if the null were true (i.e. average bolt length is 2.0 cm), is 4.8%.
  - The probability of observing data at least as extreme (i.e. an average bolt length ≤ 1.9 cm) as that which was observed, if the null were true (i.e. average bolt length is actually 2.0, is 4.8%.

Scenario 3B: A manufacturer is responsible for making barrels used to store crude oil, which are designed to hold exactly 55 gallons of oil. As part of a routine quality check, a factory manager randomly tests 27 barrels off the assembly line and finds the mean capacity is 54.7 gallons with a standard deviation of 0.8 gallons. The p-value for the appropriate test procedure was .062.

H0: the average barrel capacity is 55 gallons
Ha: the average barrel capacity is not equal to 55 gallons

- #4:
  - A nonsignificant test result (p = .062 > 0.05) means that the average barrel capacity is 55 gallons (i.e. the null is true).
  - The truth of the null hypothesis (i.e. whether the average barrel capacity is 55 gallons) is not established with a non-significant result.

- #12:
  - This p-value is properly reported as an inequality: .05 < p < .10.
  - This p-value is properly reported as an exact value: p = .062.

- #16:
  - Suppose a second random sample of 35 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .033. Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the factory manager should interpret these results as contradictory/conflicting.
  - Suppose a second random sample of 35 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .033. Despite the fact that the resulting p-values are on opposite sides of 0.05, the researcher could interpret these results as confirmatory/corroborating.

- #15:
  - Suppose a second random sample of 25 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .073. Since the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager should conclude that the totality of evidence upholds the null hypothesis (i.e. the claim that the average barrel capacity is 55 gallons).
  - Suppose a second random sample of 25 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .073. Although the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager cannot assume that the totality of evidence upholds the null hypothesis (i.e. the claim that the average barrel capacity is 55 gallons) and must consider that the studies taken collectively might lead to an entirely different conclusion than that of any individual result.

Scenario 5A: A local school district is looking at adopting a new textbook that, according to the publishers, will increase standardized test scores of second graders by more than 10 points, on average. Never willing to believe a publisher's claim without evidence to support it, the school board decides to test the claim. Two classrooms are selected for the study, one of which is assigned the new textbook and the other used the traditional text. 8 children from each class were paired based on demographics and ability level resulting in a mean difference per pair of 11.5 points with standard deviation of .76 points. The (one-tailed) p-value for the appropriate test procedure was p = .00042.

Comment [GS8]: This matching complicates things. A respondent could think that matching is imprecise and as a result disagree with all statements. Respondents may also misconstrue this study to be a t-test for independent samples since you have treatment and control classrooms. Can you make this a pre/post design?

H0: Students using the new textbook will show standardized test score gains of ≤ 10 points.
Ha: Students using the new textbook will show standardized test score gains of at least 10 points.

Comment [AD9]: If you keep this scenario, then I wouldn't use the symbol for less than or equal to and then use at least in the null and alternative. I would be consistent.

- #10:
  - The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) in this instance is 5%.
  - The probability of committing a Type I error (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) is a long-run frequency equivalent to the significance level of the test (i.e., $\alpha = 0.05$) and indeterminable for any particular hypothesis test.

- #7:
  - Statistical significance ($p = .00042 < .05$) indicates that a scientifically or substantively important relation between textbook and test scores has been detected.
  - Statistical significance ($p = .00042 < 0.05$) indicates that the observed data are unusual in relation to the null hypothesis (i.e. the new textbook yields score gains ≤ 10 points); however, the way that the data are unusual might be of no clinical interest or practical importance.

- #11:
  - If one researcher had reported the results of this study as $p < .05$ and another reported $p = .00042$, it would be correct to interpret these representations as having the same meaning.
  - Reported p-values of $p = .00042$ and $p < 0.05$ do not convey the same meaning and should not be interpreted as equivalent.

- #18:
  - Since a small p-value was observed in this study ($p = .00042$), there is a good chance that the next study testing the same hypothesis (i.e. increase in test scores ≤ 10 points) will produce a p-value at least as small.
  - Just because a small p-value was observed in this study ($p = .00042$), it cannot be assumed that the next study has a good chance to produce a p-value at least as small for the same hypothesis.

Comment [AD10]: I am not necessarily suggesting a change here, just a comment. When I am talking about hypothesis testing, I usually frame everything in terms of the alternative hypothesis because they is the hypothesis that matters and the one you make conclusions about. Here you have the question in terms of the null. You could consider changing to word in terms of the alternative.

- #14:
  - This one-sided p-value ($p = .00042$) should be converted to a two-sided value when reporting the results of the study.
  - This one-sided p-value ($p = .00042$) should be reported as-is in order to maintain compatibility with the hypothesis of practical interest.

Scenario 5B: An SAT prep course claims to increase scores by more than 60 points, on average. To test this claim, 9 students who have previously taken the SAT are randomly chosen to take the prep course. Their SAT scores before and after completing the prep course are compared and a mean difference of 81 points is found (standard deviation of 87 points). The (one-tailed) p-value for the appropriate test procedure was p = .2436.

H0: Students who take the prep course improve their SAT scores by 60 points or less.
Ha: Students who take the prep course improve their SAT scores by more than 60 points.

- #2:
    - The probability that chance produced the observed association (i.e. improved scores by 81 points for those students in the prep course) is 24.36%.
    - The probability that chance produced the observed association between SAT scores and participation in the prep course) cannot be determined from the p-value.

- #5:
    - Because this test produced a large p-value (p = .2436), this should be interpreted as evidence in favor of the claim that students who take the prep course improve their SAT score by 60 points or less
    - A large p-value cannot be said to favor the claim that the prep course leads to score improvements of ≤ 60 points (i.e. this specific null hypothesis) except in relation to those hypotheses with smaller p-values.

Comment [AD11]: Again, not sure what this means

- #6:
    - Since our p-value was greater than 0.05, this means that the study has demonstrated there to be no association between the prep course and score improvements greater than 60 points (in other words, the study has demonstrated, by virtue of the size of the p-value, that there is no evidence of a 60-point improvement effect attributable to the prep course).
    - A p-value > 0.05 (in this case, p = .2436) implies that the null hypothesis (i.e. score improvements of fewer than 60 points after prep course) is incompatible with the observed data (average improvement of 81 points); however, a conclusion of "no effect" (in other words, an effect size of zero) cannot be determined from this result.

Comment [GS12]: "…demonstrated that there is no association.…"

- #13:
    - Statistical significance is a property of the effect the prep course has on SAT scores and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.
    - Statistical significance is a property attached to the result of a statistical test and not a property of the phenomenon (i.e. the prep course's effect on SAT scores) under investigation.
- #8:
    - Lack of statistical significance (p = .2436 > .05) indicates that the effect size (i.e. the prep course's effect on SAT scores) is small.
    - The size of the effect is not determined by the significance/insignificance of the p-value; thus, a lack of statistical significance (p = .2436 > .05) does not necessarily imply that a small effect size has been detected.

Scenario 8A (ANOVA): A casino manager is concerned that one of the blackjack tables is not generating the same revenue, on average, as the other three tables. The amount of revenue received during one particular three-hour period is recorded at each of four tables each night for five consecutive nights resulting in a grand mean of $8667.90 and table means of $8486.60, $8464.20, $9509.40, and $7911.40 for Tables 1, 2, 3, & 4, respectively. The p-value for the appropriate test procedure was p = .0817.

Comment [GS13]: I suggest adding something to address ANOVA assumptions, e.g. are the dealers the same each night and don't rotate tables; are the players different each night?

H0: all tables generate the same revenue on average
Ha: there is at least one table for which the average revenue differs from the others

- #6:
  - Since our p-value was greater than 0.05 (p = .0817), this means that the study has demonstrated there to be no association between table and revenue (in other words, the study has demonstrated, by virtue of the size of the p-value, that there is no evidence of a "table" effect where at least one table is different from the others).
  - A p-value > 0.05 (in this case, p = .0817) implies that the null hypothesis (i.e. all tables generate the same revenue on average) is incompatible with the observed data (table means of $8486.60, $8464.20, $9509.40, and $7911.40); however, a conclusion of "no effect" (in other words, an effect size of zero) cannot be determined from this result.

- #8:
  - Lack of statistical significance (i.e. p = .0817 > 0.05) indicates that the effect size (effect of shift on call volume) is small.
  - The size of the effect is not determined by the significance/insignificance of the p-value; thus, a lack of statistical significance (p = .0817 > .05) does not necessarily imply that a small effect size has been detected.

- #12:
  - This p-value is properly reported as an inequality: .05 < p < .10.
  - This p-value is properly reported as an exact value: p = .0817.

- #15:
  - Suppose that in the following week, the sampling is repeated for the same three-hour period for the same four tables for five consecutive nights yielding table means of $8470.40, $8466, $9498.60, $7916.80, and $8587.95 with a p-value of p = .082. Since the same hypothesis was tested twice and neither of the tests was statistically significant, the casino manager should conclude that the totality of evidence upholds the null hypothesis (i.e. the claim that all tables generate the same revenue on average).
  - Suppose that in the following week, the sampling is repeated for the same three-hour period for the same four tables for five consecutive nights yielding table means of $8470.40, $8466, $9498.60, $7916.80, and $8587.95 with a p-value of p = .082. Although the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager cannot assume that the totality of evidence upholds the null hypothesis (i.e. the claim that all tables generate the same revenue on average) and must consider that the studies taken collectively might lead to an entirely different conclusion than that of any individual result.

Scenario 8B (ANOVA): A group of paramedics does not believe that the mean number of calls received in one shift is the same for the morning, afternoon, and evening shifts. To test this claim, they record the number of calls received during each shift for seven days, resulting in the following call/shift averages: 2.57, 3.71, and 4.28 for the morning, afternoon, and evening shifts, respectively. The p-value for the appropriate test procedure was p = .039.

H0: all shifts receive the same number of calls on average
Ha: there is at least one shift for which the average number of calls differs from the others

- #2:
  - The probability that chance produced the observed association (i.e. morning calls < afternoon calls < evening calls) is 3.9%.
  - The probability that chance produced the observed association between shift and call volume cannot be determined from the p-value.
- #13:
  - Statistical significance is a property of the effect time of day has on number of calls received and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.
  - Statistical significance is a property attached to the result of a statistical test and not a property of the phenomenon (i.e. the relationship between time of day and call volume) under investigation.
- #3:
  - A significant results here (p = .039 < 0.05) means that the null is false (i.e. that all shifts do NOT receive the same number of calls on average).
  - The falsity of the null hypothesis (i.e. all shifts receive the same number of calls on average) is not established with a significant result (i.e. p = .039 < 0.05).
- #7:
  - Statistical significance (p = .039 < 0.05) means that a substantively important relation between shift and call volume has been detected.
  - Statistical significance (p = .039 < 0.05) indicates that the observed data are unusual in relation to the null hypothesis (i.e. all shifts receive the same number of calls on average); however, the way that the data are unusual might be of no clinical interest or practical importance.

Appendix K – PV1

Point	MI#	CounterPoint
A significant test result ($p \leq 0.05$) means that the null hypothesis is false.	3	The falsity of the null hypothesis is not established with a significant result ($p \leq 0.05$)
Statistical significance indicates a scientifically or substantively important relation has been detected.	7	Statistical significance indicates that the data are unusual in relation to the null hypothesis; however, the way that the data are unusual might be of no clinical interest.
Lack of statistical significance indicates that the effect size is small.	8	Lack of statistical significance does not necessarily imply that a small effect size has been detected.
The p-value is the probability of observing the data we observed if the null hypothesis is true.	9	The p-value is the probability of observing the data we observed, combined with observations more extreme than that which was observed, if the null hypothesis is true.
When the same null hypothesis is tested in different studies and none (or a minority) of the tests are statistically significant, the totality of evidence upholds the null hypothesis.	15	When the same null hypothesis is tested in different studies, a lack of statistical significance for individual studies should not be taken as implying that the totality of evidence upholds the null hypothesis (or establishes there to be "no effect").
When the same hypothesis is tested twice, and the resulting p-values are on opposite sides of 0.05, the researcher should interpret these results as contradictory.	16	When the same hypothesis is tested twice, the fact that the resulting p-values are on opposite sides of 0.05 does not necessarily imply that these results are contradictory.
If one observes a small p-value, there is a good chance that the next study will produce a p-value at least as small for the same hypothesis.	18	If one observes a small p-value, it is incorect to assume that the next study has a good chance to produce a p-value at least as small for the same hypothesis.

False Items	MI#	True Items
The p-value is the probability that the null hypothesis is true.	1	
The p-value for the null hypothesis is the probability that chance produced the observed association.	2	
A nonsignificant test result ($p > 0.05$) means that the null hypothesis is true.	4	
A large p-value is evidence in favor of the null hypothesis.	5	A definitive conclusion (i.e. accepting a particular hypothesis) cannot be deduced from a single p-value, no matter how large.
A p-value greater than 0.05 means that the study has demonstrated there to be "no association" or "no evidence" of an effect.	6	
	8	Large effects can be masked by noise in the data and fail to achieve statistical significance, particularly in the case of small studies.
If you reject the null hypothesis because $p \leq 0.05$, the chance you have committed a Type I error (the chance your "significant finding" is a false positive) is 5%.	10	Error rates refer to the long run frequency of rejecting the hypothesis across repeated testing and not to the chance of error for any particular instance.
	11	Individual p-values ($p = 0.05$) correspond to particular observed results whereas p-value inequalities ($p < 0.05$) include a collection of possible observed results of varying degrees of incompatibility with the null hypothesis; thus, it is incorrect to interpret these results as having the same meaning.
	12	P-values should be reported as exact values unless they are so small as to fall below the numerical precision of the computation method.
	13	Statistical significance is a property attached to the result of a statistical test and not a property of the effect or population being studied.
	14	When reporting results, the nature of the p-value (i.e. one-sided vs. two-sided) must be compatible with the hypothesis of practical interest.
	16	It is possible for two studies of the same hypothesis to produce p-values on opposite sides of 0.05 despite being in perfect agreement (i.e. may show identical observed associations).
When the same hypothesis is tested twice and the resulting p-values are identical (or nearly so), the researcher should interpret these results as confirmatory.	17	It is possible for two studies of the same hypothesis to produce identical (or similar) p-values yet exhibit clearly different observed associations.

Research Vignettes			
Scenario	**MI#**	**Point**	**CounterPoint**
Scenario 2A: As indicated by a national survey, professors are irritated by students' cell phones, claiming that they ring during class an average of 15 times per semester. A reporter for the school newspaper claims that students are increasingly more courteous with their phones and that the disruptions to class have been reduced. A random sample of 12 professors found an average of 14.8 calls with a standard deviation of 1.2 calls. The (one- tailed) p-value for the appropriate test procedure was p = .288. H0: the average number of cell phone disruptions is at least 15 times per semester Ha: the average number of cell phone disruptions is less than 15 times per semester	1	There is a 28.8% chance that the number of cell phone disruptions is at least 15 times per semester (i.e. that the null is true).	The probability that the number of cell phone disruptions is at least 15 times per semester (i.e. the null is true) cannot be determined from the p-value.
	4	A nonsignificant result here (p = .288 > 0.05) means that the professors' claim (i.e. the null) is true and that the average number of cell phone disruptions is at least 15 times per semester.	The truth of the faculty's claim (i.e. the null hypothesis) is not established with a non-significant result.
	5	Because this test produced a large p-value, this should be interpreted as evidence in favor of the professors' claim (i.e. incidence of cell phone disruptions is at least 15 times per semester).	A large p-value cannot be said to favor the faculty's claim (i.e. this specific null hypothesis) except in relation to those hypotheses with smaller p-values.
	14	This one-sided p-value (p = .288) should be converted to a two-sided p-value when reporting the results of the study.	This one-sided p-value (p = .288) should be reported as- is in order to maintain compatibility with the hypothesis of practical interest.
	9	The probability of observing the data we did (i.e. an average of 14.8 calls/semester), if the null were true (i.e. average number of cell phone disruptions is actually ≥ 15 per semester), is 28.8%.	The probability of observing data at least as extreme (i.e. an average ≥ 14.8 calls/semester) as that which was observed, if the null were true (i.e. average number of cell phone disruptions is actually ≥ 15 per semester)

Scenario	MI#	Point	CounterPoint
Scenario 2B: Despite the cable company's claims of excellent customer service, most users indicate that the technicians do not report within the advertised 30-minute window. A random sample of 9 customers found an average wait of 33.2 minutes with a standard deviation of 3.4 minutes. The (one-tailed) p-value for the appropriate test procedure was p = .011. H0: the average wait time is ≤ 30 minutes Ha: the average wait time is > 30 minutes	10	The null hypothesis was rejected since p < 0.05; therefore, the chance the researcher committed a Type I error (i.e. rejecting the company's claim of wait times ≤ 30 minutes when it should not have been rejected) in this instance is 5%.	The probability of committing a Type I error (i.e. rejecting the company's claim of wait times ≤ 30 minutes when the claim is valid and should not have been rejected) is a long-run frequency equivalent to the significance level of the test (i.e., α = 0.05) and indeterminable for any particular hypothesis test.
	11	If one researcher had reported the results of this study as p < .05 and another reported p = .011, it would be correct to interpret these representations as having the same meaning.	Reported p-values of p = .011 and p < 0.05 do not convey the same meaning and should not be interpreted as equivalent.
	18	Since a small p-value was observed in this study (p = .011), there is a good chance that the next study testing the same hypothesis (i.e. customer wait times ≤ 30 minutes) will produce a p-value at least as small.	Just because a small p-value was observed in this study (p = .011), it cannot be assumed that the next study has a good chance to produce a p-value at least as small for the same hypothesis.
	16	Suppose a second random sample of 5 customers (average wait time of 33.1 minutes, standard deviation = 3.4 minutes) yielded a p-value for the appropriate test procedure of 0.056. Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the researcher should interpret these results as contradictory/conflicting.	Suppose a second random sample of 5 customers (average wait time of 33.1 minutes, standard deviation = 3.4 minutes) yielded a p-value for the appropriate test procedure of 0.056. Despite the fact that the resulting p-values are on opposite sides of 0.05, the researcher could interpret these results as confirmatory/corroborating.
	17	Suppose a second random sample of 70 customers (average wait time of 30.6 minutes, standard deviation = 2.25 minutes) yielded a p-value for the appropriate test procedure of p = 0.014. Since the same hypothesis was tested twice and the resulting p-values are nearly identical, the researcher should interpret these results as in agreement (i.e. confirmatory).	Suppose a second random sample of 70 customers (average wait time of 30.6 minutes, standard deviation = 2.25 minutes) yielded a p-value for the appropriate test procedure of p = 0.014. Despite the fact that the resulting p-values are nearly identical, the researcher should not assume these results to be in agreement.

Scenario	MI#	Point	CounterPoint
Scenario 3A: A manufacturer must test that his bolts are 2.00 cm long when they come off the assembly line. Bolts that are too long or too short cannot be used and the machines must be recalibrated in the event that unsatisfactory bolts are being produced. A random sample of 100 bolts yields a sample mean of 1.90 cm with a standard deviation of 0.5cm. The p-value for the appropriate test procedure was $p = .048$. H0: the average bolt length is 2.0 cm Ha: the average bolt length is not equal to 2.0 cm	1	There is a 4.8% chance that the average bolt length is 2.0 cm (i.e. the null is true).	The probability that the average bolt length is 2.0 cm (i.e. the null is true) cannot be determined from the p-value.
	17	Suppose a second random sample of 100 bolts (mean = 2.1 cm, standard deviation = 0.5 cm) yielded a p-value for the appropriate test procedure of $p = .048$. Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as being in agreement (i.e. confirmatory).	Suppose a second random sample of 100 bolts (mean = 2.1 cm, standard deviation = 0.5 cm) yielded a p-value for the appropriate test procedure of $p = .048$. Despite the fact that the resulting p-values are identical, it would be incorrect for the researcher to assume these results are in agreement.
	3	A significant test result ($p = .048 < .05$) means that the average bolt length is not equal to 2.0 cm (i.e. that the null is false).	The falsity of the null hypothesis (i.e. the average bolt length is not equal to 2.0 cm) is not established with a significant result (i.e. $p = .048 < 0.05$).
	9	The probability of observing the data we did (i.e. an average bolt length of 1.9 cm), if the null were true (i.e. average bolt length is 2.0 cm), is 4.8%.	The probability of observing data at least as extreme (i.e. an average bolt length $\leq$ 1.9 cm) as that which was observed, if the null were true (i.e. average bolt length is actually 2.0, is 4.8%.

Scenario	MI#	Point	CounterPoint
Scenario 3B: A manufacturer is responsible for making barrels used to store crude oil, which are designed to hold exactly 55 gallons of oil. As part of a routine quality check, a factory manager randomly tests 27 barrels off the assembly line and finds the mean capacity is 54.7 gallons with a standard deviation of 0.8 gallons. The p-value for the appropriate test procedure was .062. H0: the average barrel capacity is 55 gallons Ha: the average barrel capacity is not equal to 55 gallons	4	A nonsignificant test result (p = .062 > 0.05) means that the average barrel capacity is 55 gallons (i.e. the null is true).	The truth of the null hypothesis (i.e. whether the average barrel capacity is 55 gallons) is not established with a non-significant result.
	12	This p-value is properly reported as an inequality: .05 < p < .10.	This p-value is properly reported as an exact value: p = .062.
	16	Suppose a second random sample of 35 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .033. Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the factory manager should interpret these results as contradictory/conflicting.	Suppose a second random sample of 35 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .033. Despite the fact that the resulting p-values are on opposite sides of 0.05, the researcher could interpret these results as confirmatory/corroborating.
	15	Suppose a second random sample of 25 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .073. Since the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager should conclude that the totality of evidence upholds the null hypothesis (i.e. the claim that the average barrel capacity is 55 gallons).	Suppose a second random sample of 25 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .073. Although the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager cannot assume that the totality of evidence upholds the null hypothesis (i.e. the claim that the average barrel capacity is 55 gallons) and must consider that the studies taken collectively might lead to an entirely different conclusion than that of any individual result.

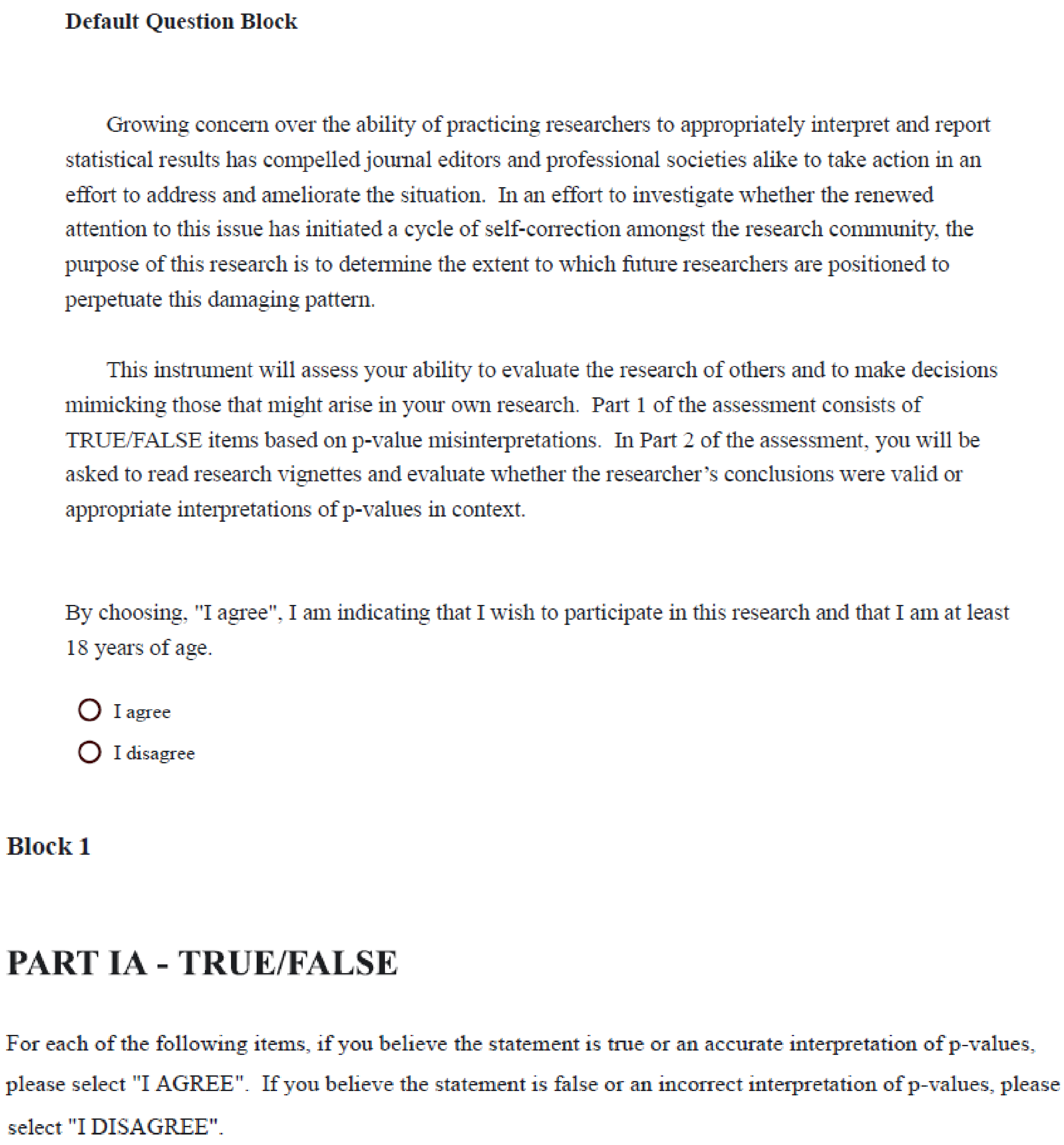
Scenario	MI#	Point	CounterPoint
Scenario 5A: A local school district is looking at adopting a new textbook that, according to the publishers, will increase standardized test scores of second graders by more than 10 points, on average. Never willing to believe a publisher's claim without evidence to support it, the school board decides to test the claim. Two classrooms are selected for the study, one of which is assigned the new textbook and the other used the traditional text. 8 children from each class were paired based on demographics and ability level resulting in a mean difference per pair of 11.5 points with standard deviation of .76 points. The (one-tailed) p-value for the appropriate test procedure was $p = .00042$. You may assume that all underlying assumptions have been satisfied and that there are no issues related to the experimental design (e.g. imprecise matching). H0: The difference between the groups will be ≤ 10 points. Ha: The difference between the groups will be > 10 points.	10	The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) in this instance is 5%.	The probability of committing a Type I error (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) is a long-run frequency equivalent to the significance level of the test (i.e., $\alpha = 0.05$) and indeterminable for any particular hypothesis test.
	7	Statistical significance ($p = .00042 < .05$) indicates that a scientifically or substantively important relation between textbook and test scores has been detected.	Statistical significance ($p = .00042 < 0.05$) indicates that the observed data are unusual in relation to the null hypothesis (i.e. the new textbook yields score gains ≤ 10 points); however, the way that the data are unusual might be of no clinical interest or practical importance.
	11	If one researcher had reported the results of this study as $p < .05$ and another reported $p = .00042$, it would be correct to interpret these representations as having the same meaning.	Reported p-values of $p = .00042$ and $p < 0.05$ do not convey the same meaning and should not be interpreted as equivalent.
	18	Since a small p-value was observed in this study ($p = .00042$), there is a good chance that the next study testing the same hypothesis (i.e. increase in test scores > 10 points) will produce a p-value at least as small.	Just because a small p-value was observed in this study ($p = .00042$), it cannot be assumed that the next study has a good chance to produce a p-value at least as small for the same hypothesis.
	14	This one-sided p-value ($p = .00042$) should be converted to a two-sided value when reporting the results of the study.	This one-sided p-value ($p = .00042$) should be reported as-is in order to maintain compatibility with the hypothesis of practical interest.

Scenario	MI#	Point	CounterPoint
Scenario 5B: An SAT prep course claims to increase scores by more than 60 points, on average. To test this claim, 9 students who have previously taken the SAT are randomly chosen to take the prep course. Their SAT scores before and after completing the prep course are compared and a mean difference of 81 points is found (standard deviation of 87 points). The (one-tailed) p-value for the appropriate test procedure was p = .2436. H0: Students who take the prep course improve their SAT scores by 60 points or less. Ha: Students who take the prep course improve their SAT scores by more than 60 points.	2	The probability that chance produced the observed association (i.e. improved scores by 81 points for those students in the prep course) is 24.36%.	The probability that chance produced the observed association between SAT scores and participation in the prep course) cannot be determined from the p-value.
	5	Because this test produced a large p-value (p = .2436), this should be interpreted as evidence in favor of the claim that students who take the prep course improve their SAT score by 60 points or less	A large p-value cannot be said to favor the claim that the prep course leads to score improvements of ≤ 60 points (i.e. this specific null hypothesis) except in relation to those hypotheses with smaller p-values.
	6	Since our p-value was greater than 0.05, this means that the study has demonstrated there is no association between the prep course and score improvements greater than 60 points (in other words, the study has demonstrated, by virtue of the size of the p-value, that there is no evidence of a 60-point improvement effect attributable to the prep course).	A p-value > 0.05 (in this case, p = .2436) implies that the null hypothesis (i.e. score improvements of fewer than 60 points after prep course) is incompatible with the observed data (average improvement of 81 points); however, a conclusion of "no effect" (in other words, an effect size of zero) cannot be determined from this result.
	13	Statistical significance is a property of the effect the prep course has on SAT scores and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.	Statistical significance is a property attached to the result of a statistical test and not a property of the phenomenon (i.e. the prep course's effect on SAT scores) under investigation.
	8	Lack of statistical significance (p = .2436 > .05) indicates that the effect size (i.e. the prep course's effect on SAT scores) is small.	The size of the effect is not determined by the significance/insignificance of the p-value; thus, a lack of statistical significance (p = .2436 > .05) does not necessarily imply that a small effect size has been detected.

Scenario	MI#	Point	CounterPoint
Scenario 8A (ANOVA): A casino manager is concerned that one of the blackjack tables is not generating the same revenue, on average, as the other three tables. The amount of revenue received during one particular three-hour period is recorded at each of four tables each night for five consecutive nights resulting in a grand mean of $8667.90 and table means of $8486.60, $8464.20, $9509.40, and $7911.40 for Tables 1, 2, 3, & 4, respectively. The p-value for the appropriate test procedure was p = .0817. You may assume that all underlying assumptions have been satisfied and that there are no issues related to experimental design (e.g., dealer/player swaps/rotations). H0: all tables generate the same revenue on average Ha: there is at least one table for which the average revenue differs from the others	6	Since our p-value was greater than 0.05 (p = .0817), this means that the study has demonstrated there to be no association between table and revenue (in other words, the study has demonstrated, by virtue of the size of the p-value, that there is no evidence of a "table" effect where at least one table is different from the others).	A p-value > 0.05 (in this case, p = .0817) implies that the null hypothesis (i.e. all tables generate the same revenue on average) is incompatible with the observed data (table means of $8486.60, $8464.20, $9509.40, and $7911.40); however, a conclusion of "no effect" (in other words, an effect size of zero) cannot be determined from this result.
	8	Lack of statistical significance (i.e. p = .0817 > 0.05) indicates that the effect size (effect of shift on call volume) is small.	The size of the effect is not determined by the significance/insignificance of the p-value; thus, a lack of statistical significance (p = .0817 > .05) does not necessarily imply that a small effect size has been detected.
	12	This p-value is properly reported as an inequality: .05 < p < .10.	This p-value is properly reported as an exact value: p = .0817.
	15	Suppose that in the following week, the sampling is repeated for the same three-hour period for the same four tables for five consecutive nights yielding table means of $8470.40, $8466, $9498.60, $7916.80, and $8587.95 with a p-value of p = .082. Since the same hypothesis was tested twice and neither of the tests was statistically significant, the casino manager should conclude that the totality of evidence upholds the null hypothesis (i.e. the claim that all tables generate the same revenue on average).	Suppose that in the following week, the sampling is repeated for the same three-hour period for the same four tables for five consecutive nights yielding table means of $8470.40, $8466, $9498.60, $7916.80, and $8587.95 with a p-value of p = .082. Although the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager cannot assume that the totality of evidence upholds the null hypothesis (i.e. the claim that all tables generate the same revenue on average) and must consider that the studies taken collectively might lead to an entirely different conclusion than that of any individual result.

Scenario	MI#	Point	CounterPoint
Scenario 8B (ANOVA): A group of paramedics does not believe that the mean number of calls received in one shift is the same for the morning, afternoon, and evening shifts. To test this claim, they record the number of calls received during each shift for seven days, resulting in the following call/shift averages: 2.57, 3.71, and 4.28 for the morning, afternoon, and evening shifts, respectively. The p-value for the appropriate test procedure was p = .039. H0: all shifts receive the same number of calls on average Ha: there is at least one shift for which the average number of calls differs from the others	2	The probability that chance produced the observed association (i.e. morning calls < afternoon calls < evening calls) is 3.9%.	The probability that chance produced the observed association between shift and call volume cannot be determined from the p-value.
	13	Statistical significance is a property of the effect time of day has on number of calls received and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.	Statistical significance is a property attached to the result of a statistical test and not a property of the phenomenon (i.e. the relationship between time of day and call volume) under investigation.
	3	A significant results here (p = .039 < 0.05) means that the null is false (i.e. that all shifts do NOT receive the same number of calls on average).	The falsity of the null hypothesis (i.e. all shifts receive the same number of calls on average) is not established with a significant result (i.e. p = .039 < 0.05).
	7	Statistical significance (p = .039 < 0.05) means that a substantively important relation between shift and call volume has been detected.	Statistical significance (p = .039 < 0.05) indicates that the observed data are unusual in relation to the null hypothesis (i.e. all shifts receive the same number of calls on average); however, the way that the data are unusual might be of no clinical interest or practical importance.

Appendix L – PV1 Qualtrics Version (Phase IIA&B version)



Default Question Block

Growing concern over the ability of practicing researchers to appropriately interpret and report statistical results has compelled journal editors and professional societies alike to take action in an effort to address and ameliorate the situation. In an effort to investigate whether the renewed attention to this issue has initiated a cycle of self-correction amongst the research community, the purpose of this research is to determine the extent to which future researchers are positioned to perpetuate this damaging pattern.

This instrument will assess your ability to evaluate the research of others and to make decisions mimicking those that might arise in your own research. Part 1 of the assessment consists of TRUE/FALSE items based on p-value misinterpretations. In Part 2 of the assessment, you will be asked to read research vignettes and evaluate whether the researcher's conclusions were valid or appropriate interpretations of p-values in context.

By choosing, "I agree", I am indicating that I wish to participate in this research and that I am at least 18 years of age.

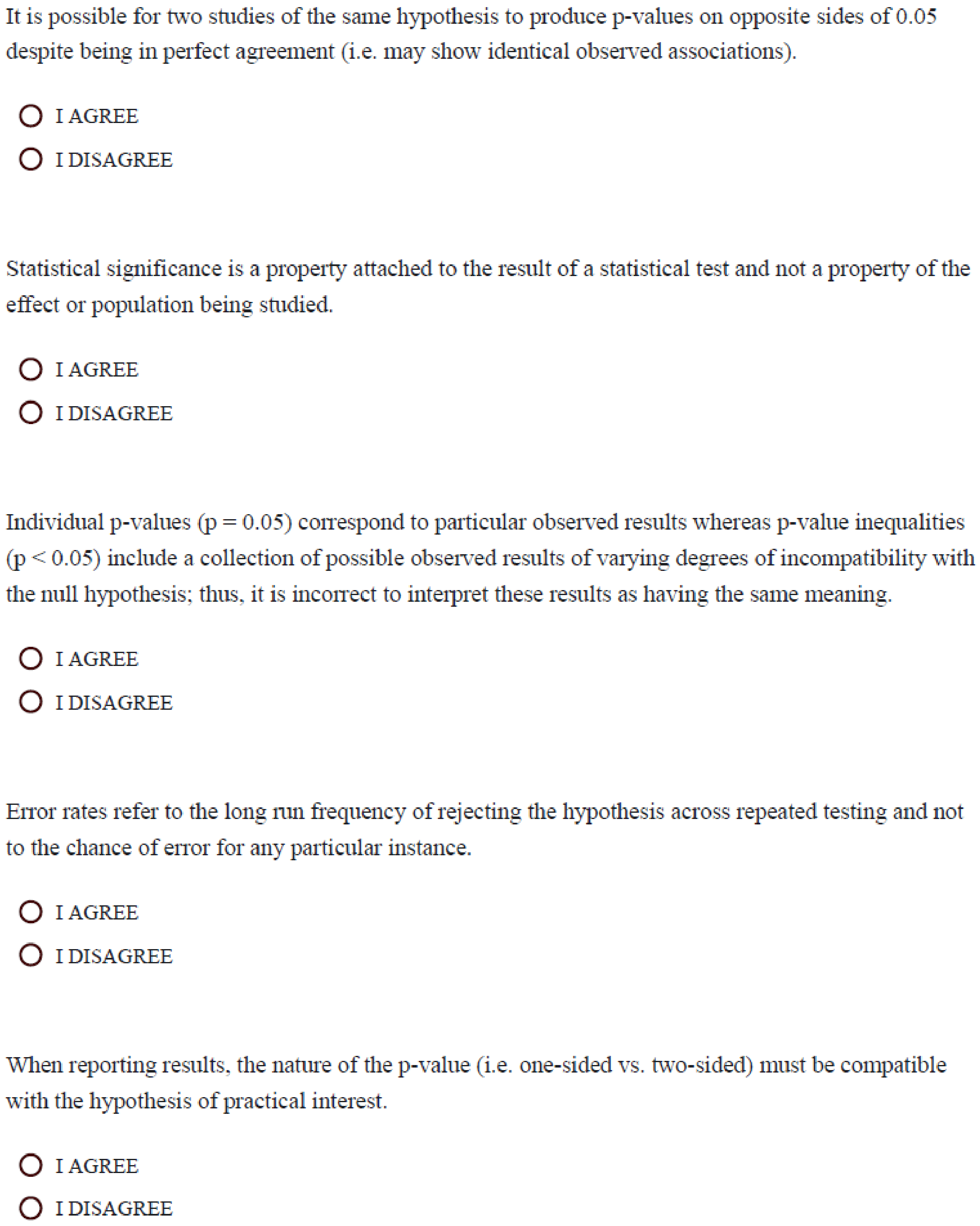
○ I agree

○ I disagree

Block 1

PART IA - TRUE/FALSE

For each of the following items, if you believe the statement is true or an accurate interpretation of p-values, please select "I AGREE". If you believe the statement is false or an incorrect interpretation of p-values, please select "I DISAGREE".



It is possible for two studies of the same hypothesis to produce p-values on opposite sides of 0.05 despite being in perfect agreement (i.e. may show identical observed associations).

○ I AGREE

○ I DISAGREE

Statistical significance is a property attached to the result of a statistical test and not a property of the effect or population being studied.

○ I AGREE

○ I DISAGREE

Individual p-values (p = 0.05) correspond to particular observed results whereas p-value inequalities (p < 0.05) include a collection of possible observed results of varying degrees of incompatibility with the null hypothesis; thus, it is incorrect to interpret these results as having the same meaning.

○ I AGREE

○ I DISAGREE

Error rates refer to the long run frequency of rejecting the hypothesis across repeated testing and not to the chance of error for any particular instance.

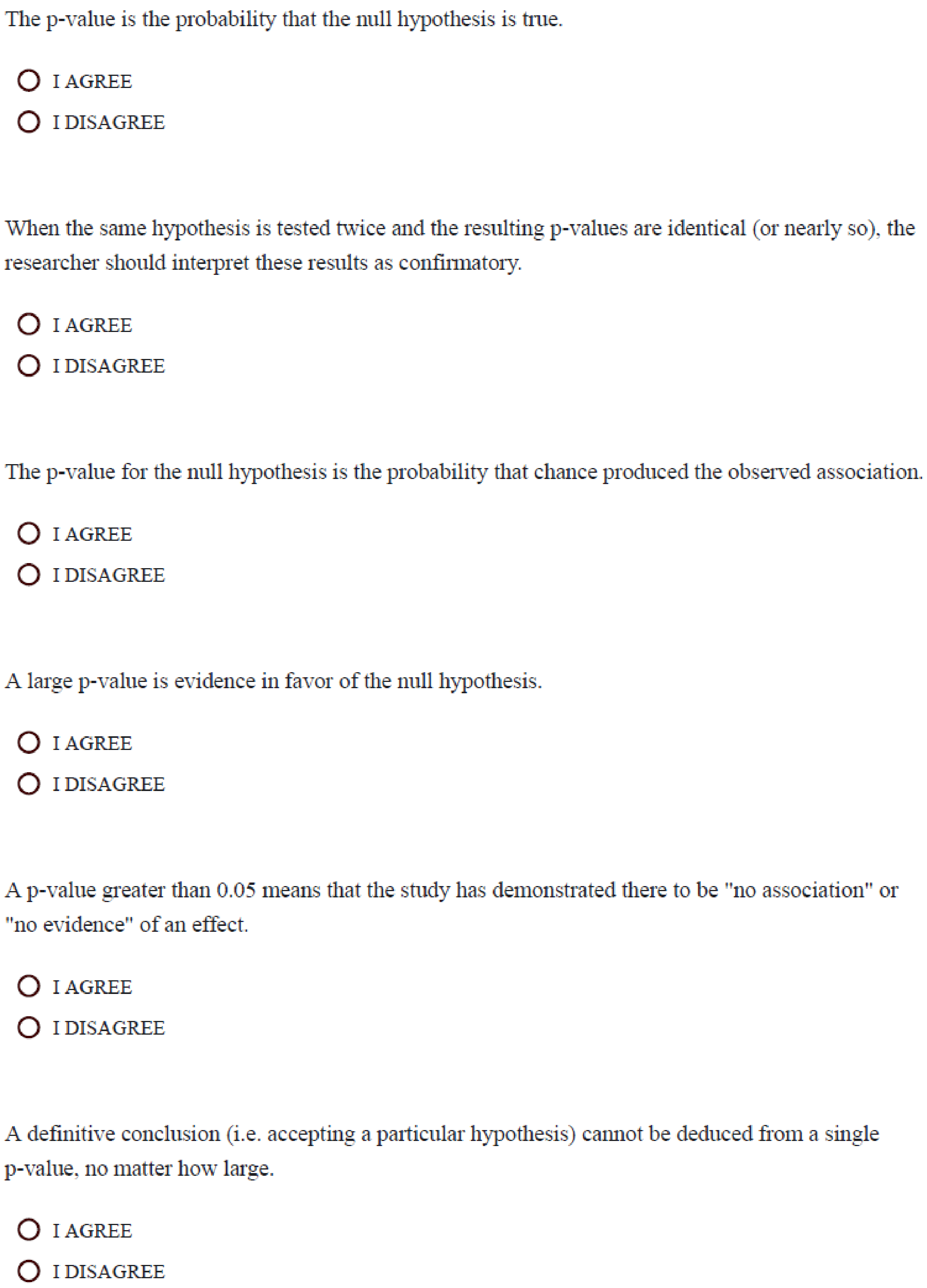
○ I AGREE

○ I DISAGREE

When reporting results, the nature of the p-value (i.e. one-sided vs. two-sided) must be compatible with the hypothesis of practical interest.

○ I AGREE

○ I DISAGREE



The p-value is the probability that the null hypothesis is true.

○ I AGREE

○ I DISAGREE

When the same hypothesis is tested twice and the resulting p-values are identical (or nearly so), the researcher should interpret these results as confirmatory.

○ I AGREE

○ I DISAGREE

The p-value for the null hypothesis is the probability that chance produced the observed association.

○ I AGREE

○ I DISAGREE

A large p-value is evidence in favor of the null hypothesis.

○ I AGREE

○ I DISAGREE

A p-value greater than 0.05 means that the study has demonstrated there to be "no association" or "no evidence" of an effect.

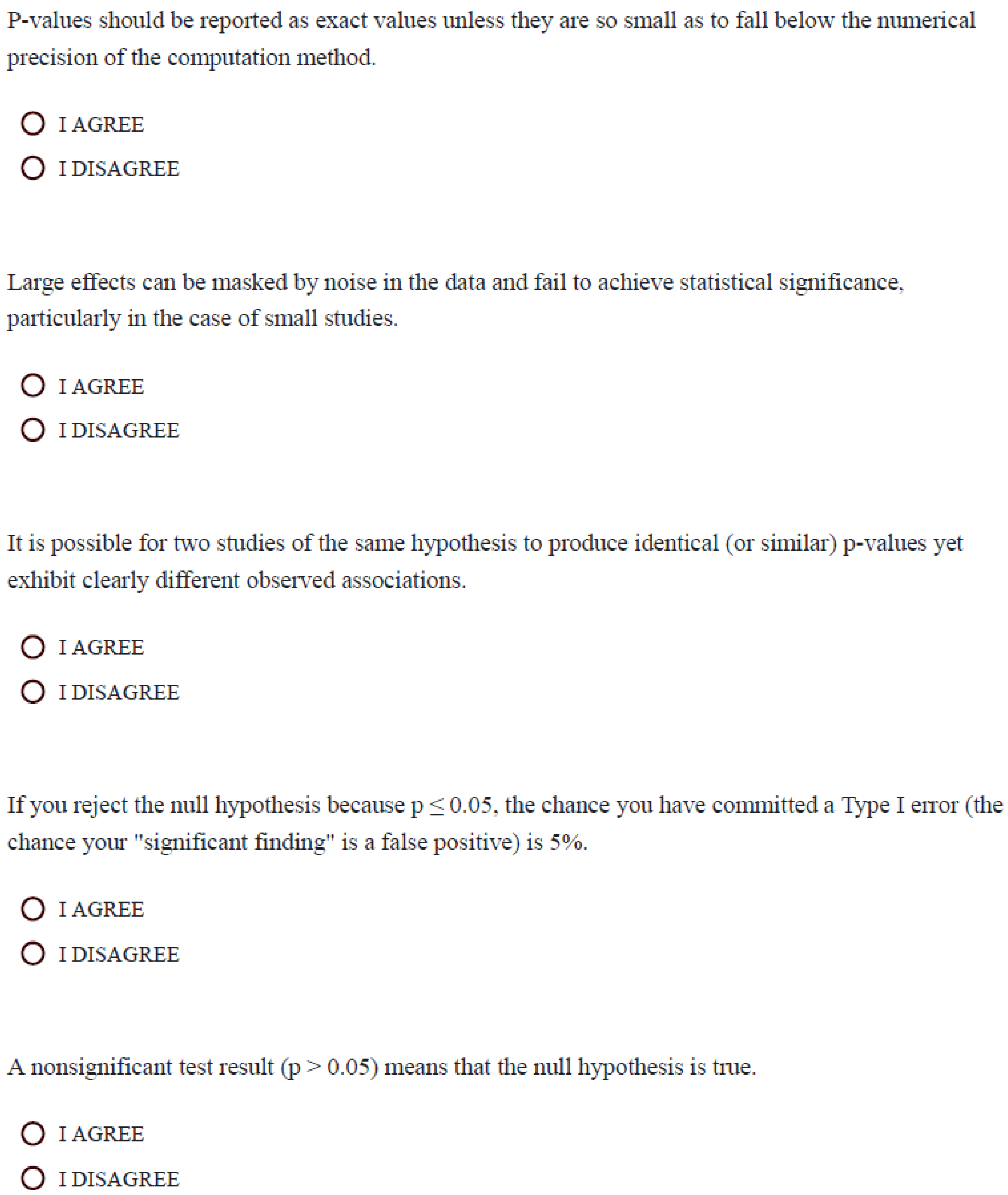
○ I AGREE

○ I DISAGREE

A definitive conclusion (i.e. accepting a particular hypothesis) cannot be deduced from a single p-value, no matter how large.

○ I AGREE

○ I DISAGREE



P-values should be reported as exact values unless they are so small as to fall below the numerical precision of the computation method.

○ I AGREE

○ I DISAGREE

Large effects can be masked by noise in the data and fail to achieve statistical significance, particularly in the case of small studies.

○ I AGREE

○ I DISAGREE

It is possible for two studies of the same hypothesis to produce identical (or similar) p-values yet exhibit clearly different observed associations.

○ I AGREE

○ I DISAGREE

If you reject the null hypothesis because $p \leq 0.05$, the chance you have committed a Type I error (the chance your "significant finding" is a false positive) is 5%.

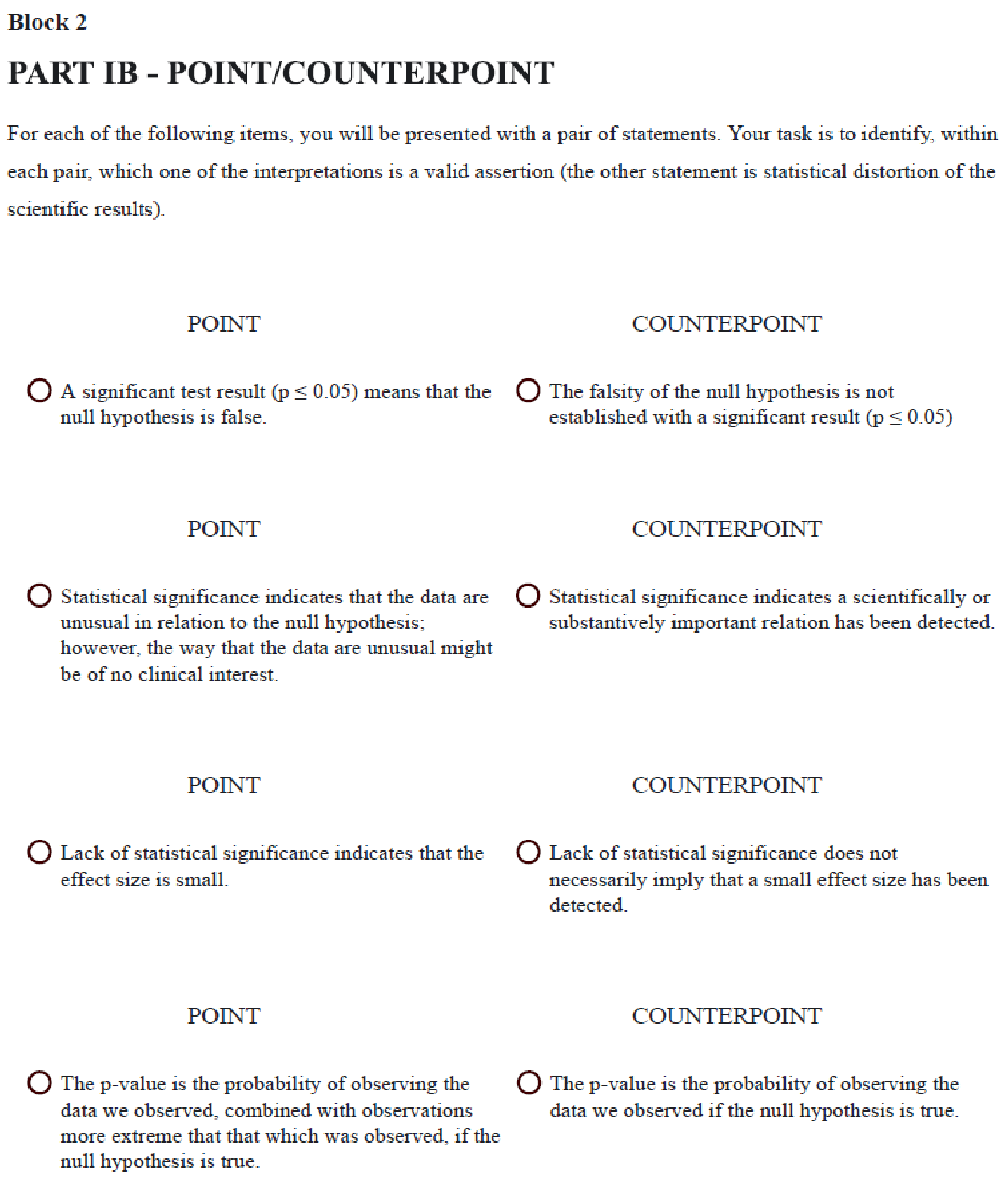
○ I AGREE

○ I DISAGREE

A nonsignificant test result ($p > 0.05$) means that the null hypothesis is true.

○ I AGREE

○ I DISAGREE



Block 2

PART IB - POINT/COUNTERPOINT

For each of the following items, you will be presented with a pair of statements. Your task is to identify, within each pair, which one of the interpretations is a valid assertion (the other statement is statistical distortion of the scientific results).

POINT	COUNTERPOINT
○ A significant test result ($p \leq 0.05$) means that the null hypothesis is false.	○ The falsity of the null hypothesis is not established with a significant result ($p \leq 0.05$)

POINT	COUNTERPOINT
○ Statistical significance indicates that the data are unusual in relation to the null hypothesis; however, the way that the data are unusual might be of no clinical interest.	○ Statistical significance indicates a scientifically or substantively important relation has been detected.

POINT	COUNTERPOINT
○ Lack of statistical significance indicates that the effect size is small.	○ Lack of statistical significance does not necessarily imply that a small effect size has been detected.

POINT	COUNTERPOINT
○ The p-value is the probability of observing the data we observed, combined with observations more extreme that that which was observed, if the null hypothesis is true.	○ The p-value is the probability of observing the data we observed if the null hypothesis is true.

POINT

○ When the same null hypothesis is tested in different studies and none (or a minority) of the tests are statistically significant, the totality of evidence upholds the null hypothesis.

COUNTERPOINT

○ When the same null hypothesis is tested in different studies, a lack of statistical significance for individual studies should not be taken as implying that the totality of evidence upholds the null hypothesis (or establishes there to be "no effect").

POINT

○ When the same hypothesis is tested twice, the fact that the resulting p-values are on opposite sides of 0.05 does not necessarily imply that these results are contradictory.

COUNTERPOINT

○ When the same hypothesis is tested twice, and the resulting p-values are on opposite sides of 0.05, the researcher should interpret these results as contradictory.

POINT

○ If one observes a small p-value, there is a good chance that the next study will produce a p-value at least as small for the same hypothesis.

COUNTERPOINT

○ If one observes a small p-value, it is incorect to assume that the next study has a good chance to produce a p-value at least as small for the same hypothesis.

Block 3

Part II - Research Vignettes

In the following section, you will be presented with a collection of research vignettes, each of which is accompanied by a series of paired statements meant to represent hypothetical interpretations of the research results. In the spirit of peer review, your task is to identify, within each pair, which one of the interpretations is a valid assertion (the other statement is statistical distortion of the scientific results).

Scenario 2A: As indicated by a national survey, professors are irritated by students' cell phones, claiming that they ring during class an average of 15 times per semester. A reporter for the school newspaper claims that students are increasingly more courteous with their phones and that the disruptions to class have been reduced. A random sample of 12 professors found an average of 14.8 calls with a standard deviation of 1.2 calls. The (one-tailed) p-value for the appropriate test procedure was p = .288.

H0: the average number of cell phone disruptions is at least 15 times per semester
Ha: the average number of cell phone disruptions is less than 15 times per semester

2A1

○ There is a 28.8% chance that the number of cell phone disruptions is at least 15 times per semester (i.e. that the null is true).

○ The probability that the number of cell phone disruptions is at least 15 times per semester (i.e. the null is true) cannot be determined from the p-value.

2A4

○ The truth of the faculty's claim (i.e. the null hypothesis) is not established with a non-significant result.

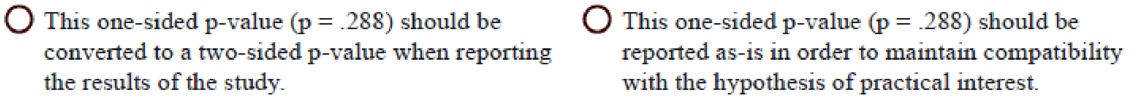
○ A nonsignificant result here (p = .288 > 0.05) means that the professors' claim (i.e. the null) is true and that the average number of cell phone disruptions is at least 15 times per semester.

2A5

○ Because this test produced a large p-value, this should be interpreted as evidence in favor of the professors' claim (i.e. incidence of cell phone disruptions is at least 15 times per semester).

○ A large p-value cannot be said to favor the faculty's claim (i.e. this specific null hypothesis) except in relation to those hypotheses with smaller p-values.

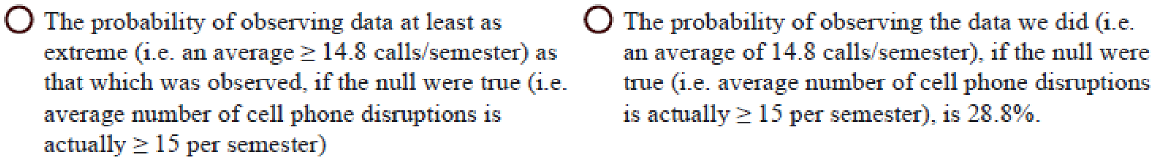
2A14



○ This one-sided p-value (p = .288) should be converted to a two-sided p-value when reporting the results of the study.

○ This one-sided p-value (p = .288) should be reported as-is in order to maintain compatibility with the hypothesis of practical interest.

2A9



○ The probability of observing data at least as extreme (i.e. an average ≥ 14.8 calls/semester) as that which was observed, if the null were true (i.e. average number of cell phone disruptions is actually ≥ 15 per semester)

○ The probability of observing the data we did (i.e. an average of 14.8 calls/semester), if the null were true (i.e. average number of cell phone disruptions is actually ≥ 15 per semester), is 28.8%.

Block 4

Scenario 2B: Despite the cable company's claims of excellent customer service, most users indicate that the technicians do not report within the advertised 30-minute window. A random sample of 9 customers found an average wait of 33.2 minutes with a standard deviation of 3.4 minutes. The (one-tailed) p-value for the appropriate test procedure was p = .011.

H0: the average wait time is ≤ 30 minutes
Ha: the average wait time is > 30 minutes

2B10

○ The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error (i.e. rejecting the company's claim of wait times ≤ 30 minutes when it should not have been rejected) in this instance is 5%.

○ The probability of committing a Type I error (i.e. rejecting the company's claim of wait times ≤ 30 minutes when the claim is valid and should not have been rejected) is a long-run frequency equivalent to the significance level of the test (i.e., $\alpha = 0.05$) and indeterminable for any particular hypothesis test.

2B11

○ Reported p-values of $p = .011$ and $p < 0.05$ do not convey the same meaning and should not be interpreted as equivalent.

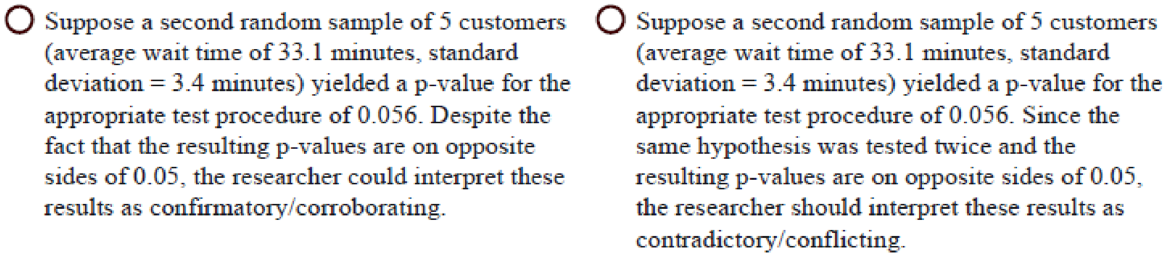
○ If one researcher had reported the results of this study as $p < .05$ and another reported $p = .011$, it would be correct to interpret these representations as having the same meaning.

2B18

○ Since a small p-value was observed in this study ($p = .011$), there is a good chance that the next study testing the same hypothesis (i.e. customer wait times ≤ 30 minutes) will produce a p-value at least as small.

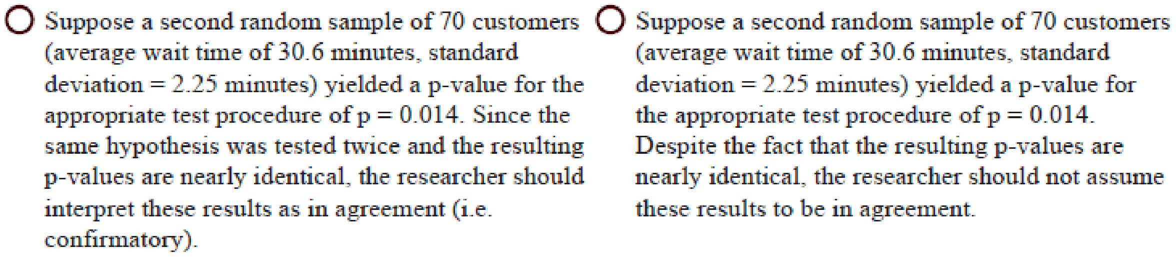
○ Just because a small p-value was observed in this study ($p = .011$), it cannot be assumed that the next study has a good chance to produce a p-value at least as small for the same hypothesis.

2B16


○ Suppose a second random sample of 5 customers (average wait time of 33.1 minutes, standard deviation = 3.4 minutes) yielded a p-value for the appropriate test procedure of 0.056. Despite the fact that the resulting p-values are on opposite sides of 0.05, the researcher could interpret these results as confirmatory/corroborating.

○ Suppose a second random sample of 5 customers (average wait time of 33.1 minutes, standard deviation = 3.4 minutes) yielded a p-value for the appropriate test procedure of 0.056. Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the researcher should interpret these results as contradictory/conflicting.

2B17


○ Suppose a second random sample of 70 customers (average wait time of 30.6 minutes, standard deviation = 2.25 minutes) yielded a p-value for the appropriate test procedure of p = 0.014. Since the same hypothesis was tested twice and the resulting p-values are nearly identical, the researcher should interpret these results as in agreement (i.e. confirmatory).

○ Suppose a second random sample of 70 customers (average wait time of 30.6 minutes, standard deviation = 2.25 minutes) yielded a p-value for the appropriate test procedure of p = 0.014. Despite the fact that the resulting p-values are nearly identical, the researcher should not assume these results to be in agreement.

Block 5

Scenario 3A: A manufacturer must test that his bolts are 2.00 cm long when they come off the assembly line. Bolts that are too long or too short cannot be used and the machines must be recalibrated in the event that unsatisfactory bolts are being produced. A random sample of 100 bolts yields a sample mean of 1.90 cm with a standard deviation of 0.5cm. The p-value for the appropriate test procedure was $p = .048$.

H0: the average bolt length is 2.0 cm
Ha: the average bolt length is not equal to 2.0 cm

3A1

○ There is a 4.8% chance that the average bolt length is 2.0 cm (i.e. the null is true).

○ The probability that the average bolt length is 2.0 cm (i.e. the null is true) cannot be determined from the p-value.

3A17

○ Suppose a second random sample of 100 bolts (mean = 2.1 cm, standard deviation = 0.5 cm) yielded a p-value for the appropriate test procedure of $p = .048$. Despite the fact that the resulting p-values are identical, it would be incorrect for the researcher to assume these results are in agreement.

○ Suppose a second random sample of 100 bolts (mean = 2.1 cm, standard deviation = 0.5 cm) yielded a p-value for the appropriate test procedure of $p = .048$. Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as being in agreement (i.e. confirmatory).

3A3

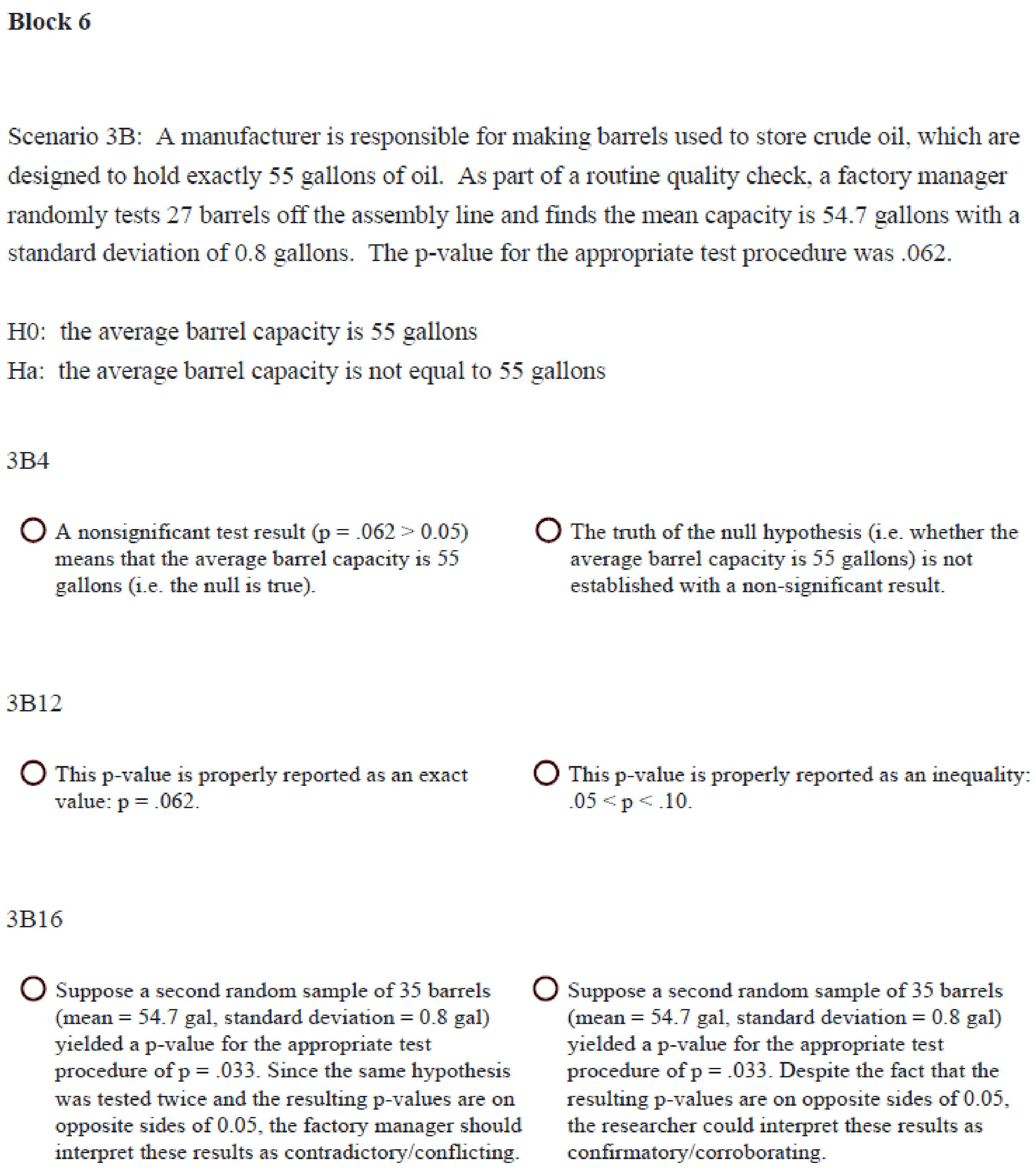
○ A significant test result ($p = .048 < .05$) means that the average bolt length is not equal to 2.0 cm (i.e. that the null is false).

○ The falsity of the null hypothesis (i.e. the average bolt length is not equal to 2.0 cm) is not established with a significant result (i.e. $p = .048 < 0.05$).

3A9

○ The probability of observing data at least as extreme (i.e. an average bolt length ≤ 1.9 cm) as that which was observed, if the null were true (i.e. average bolt length is actually 2.0, is 4.8%.

○ The probability of observing the data we did (i.e. an average bolt length of 1.9 cm), if the null were true (i.e. average bolt length is 2.0 cm), is 4.8%.



Block 6

Scenario 3B: A manufacturer is responsible for making barrels used to store crude oil, which are designed to hold exactly 55 gallons of oil. As part of a routine quality check, a factory manager randomly tests 27 barrels off the assembly line and finds the mean capacity is 54.7 gallons with a standard deviation of 0.8 gallons. The p-value for the appropriate test procedure was .062.

H0: the average barrel capacity is 55 gallons
Ha: the average barrel capacity is not equal to 55 gallons

3B4

○ A nonsignificant test result (p = .062 > 0.05) means that the average barrel capacity is 55 gallons (i.e. the null is true).

○ The truth of the null hypothesis (i.e. whether the average barrel capacity is 55 gallons) is not established with a non-significant result.

3B12

○ This p-value is properly reported as an exact value: p = .062.

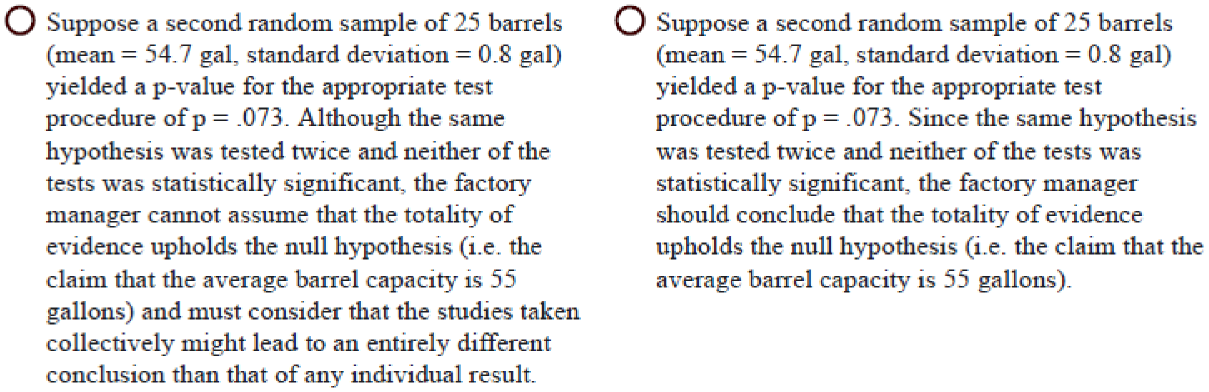
○ This p-value is properly reported as an inequality: .05 < p < .10.

3B16

○ Suppose a second random sample of 35 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .033. Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the factory manager should interpret these results as contradictory/conflicting.

○ Suppose a second random sample of 35 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .033. Despite the fact that the resulting p-values are on opposite sides of 0.05, the researcher could interpret these results as confirmatory/corroborating.

3B15



○ Suppose a second random sample of 25 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .073. Although the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager cannot assume that the totality of evidence upholds the null hypothesis (i.e. the claim that the average barrel capacity is 55 gallons) and must consider that the studies taken collectively might lead to an entirely different conclusion than that of any individual result.

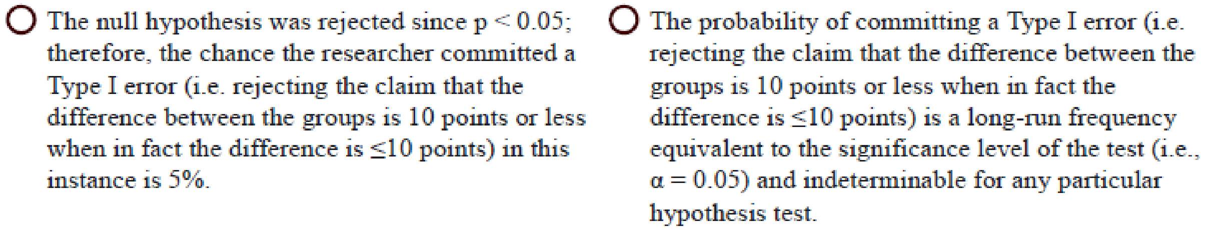
○ Suppose a second random sample of 25 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .073. Since the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager should conclude that the totality of evidence upholds the null hypothesis (i.e. the claim that the average barrel capacity is 55 gallons).

Block 7

Scenario 5A: A local school district is looking at adopting a new textbook that, according to the publishers, will increase standardized test scores of second graders by more than 10 points, on average. Never willing to believe a publisher's claim without evidence to support it, the school board decides to test the claim. Two classrooms are selected for the study, one of which is assigned the new textbook and the other used the traditional text. 8 children from each class were paired based on demographics and ability level resulting in a mean difference per pair of 11.5 points with standard deviation of .76 points. The (one-tailed) p-value for the appropriate test procedure was $p = .00042$. You may assume that all underlying assumptions have been satisfied and that there are no issues related to the experimental design (e.g. imprecise matching).

H0: The difference between the groups will be ≤ 10 points.
Ha: The difference between the groups will be > 10 points.

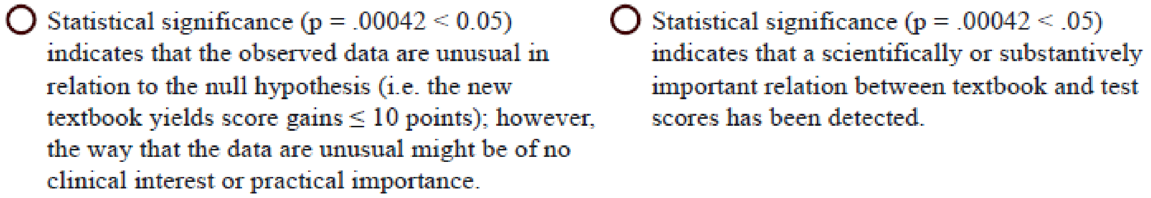
5A10



○ The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) in this instance is 5%.

○ The probability of committing a Type I error (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) is a long-run frequency equivalent to the significance level of the test (i.e., $\alpha = 0.05$) and indeterminable for any particular hypothesis test.

5A7



○ Statistical significance ($p = .00042 < 0.05$) indicates that the observed data are unusual in relation to the null hypothesis (i.e. the new textbook yields score gains ≤ 10 points); however, the way that the data are unusual might be of no clinical interest or practical importance.

○ Statistical significance ($p = .00042 < .05$) indicates that a scientifically or substantively important relation between textbook and test scores has been detected.

5A11

○ If one researcher had reported the results of this study as $p < .05$ and another reported $p = .00042$, it would be correct to interpret these representations as having the same meaning.

○ Reported p-values of $p = .00042$ and $p < 0.05$ do not convey the same meaning and should not be interpreted as equivalent.

5A18

○ Just because a small p-value was observed in this study ($p = .00042$), it cannot be assumed that the next study has a good chance to produce a p-value at least as small for the same hypothesis.

○ Since a small p-value was observed in this study ($p = .00042$), there is a good chance that the next study testing the same hypothesis (i.e. increase in test scores > 10 points) will produce a p-value at least as small.

5A14

○ This one-sided p-value ($p = .00042$) should be converted to a two-sided value when reporting the results of the study.

○ This one-sided p-value ($p = .00042$) should be reported as-is in order to maintain compatibility with the hypothesis of practical interest.

Block 8

Scenario 5B: An SAT prep course claims to increase scores by more than 60 points, on average. To test this claim, 9 students who have previously taken the SAT are randomly chosen to take the prep course. Their SAT scores before and after completing the prep course are compared and a mean difference of 81 points is found (standard deviation of 87 points). The (one-tailed) p-value for the appropriate test procedure was p = .2436.

H0: Students who take the prep course improve their SAT scores by 60 points or less.
Ha: Students who take the prep course improve their SAT scores by more than 60 points.

5B2

○ The probability that chance produced the observed association (i.e. improved scores by 81 points for those students in the prep course) is 24.36%.

○ The probability that chance produced the observed association between SAT scores and participation in the prep course) cannot be determined from the p-value.

5B5

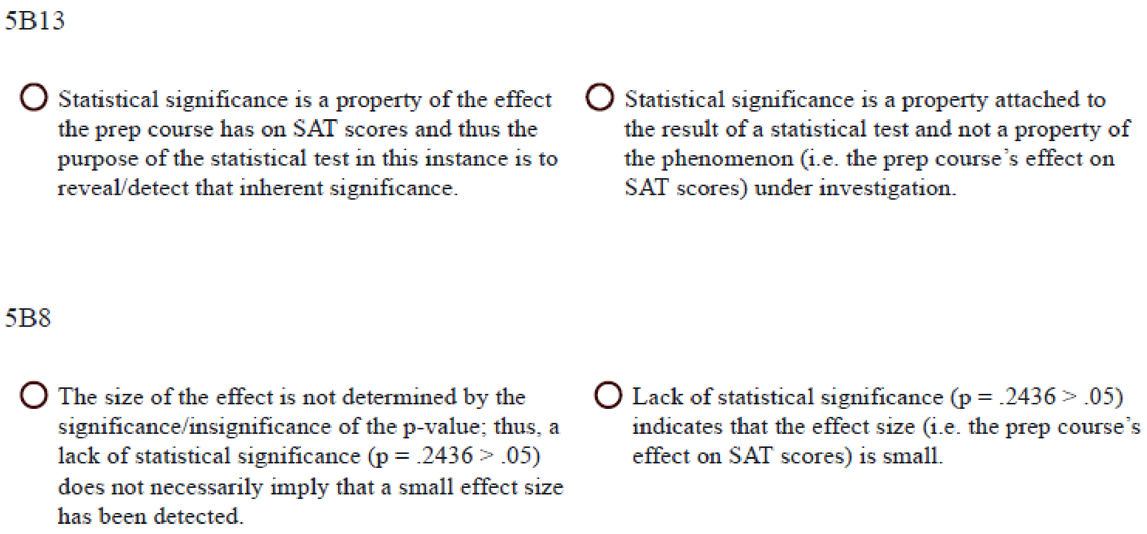
○ A large p-value cannot be said to favor the claim that the prep course leads to score improvements of ≤ 60 points (i.e. this specific null hypothesis) except in relation to those hypotheses with smaller p-values.

○ Because this test produced a large p-value (p = .2436), this should be interpreted as evidence in favor of the claim that students who take the prep course improve their SAT score by 60 points or less

5B6

○ Since our p-value was greater than 0.05, this means that the study has demonstrated there is no association between the prep course and score improvements greater than 60 points (in other words, the study has demonstrated, by virtue of the size of the p-value, that there is no evidence of a 60-point improvement effect attributable to the prep course).

○ A p-value > 0.05 (in this case, p = .2436) implies that the null hypothesis (i.e. score improvements of fewer than 60 points after prep course) is incompatible with the observed data (average improvement of 81 points); however, a conclusion of "no effect" (in other words, an effect size of zero) cannot be determined from this result.



5B13

○ Statistical significance is a property of the effect the prep course has on SAT scores and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.

○ Statistical significance is a property attached to the result of a statistical test and not a property of the phenomenon (i.e. the prep course's effect on SAT scores) under investigation.

5B8

○ The size of the effect is not determined by the significance/insignificance of the p-value; thus, a lack of statistical significance (p = .2436 > .05) does not necessarily imply that a small effect size has been detected.

○ Lack of statistical significance (p = .2436 > .05) indicates that the effect size (i.e. the prep course's effect on SAT scores) is small.

Block 9

Scenario 8A (ANOVA): A casino manager is concerned that one of the blackjack tables is not generating the same revenue, on average, as the other three tables. The amount of revenue received during one particular three-hour period is recorded at each of four tables each night for five consecutive nights resulting in a grand mean of $8667.90 and table means of $8486.60, $8464.20, $9509.40, and $7911.40 for Tables 1, 2, 3, & 4, respectively. The p-value for the appropriate test procedure was p = .0817. You may assume that all underlying assumptions have been satisfied and that there are no issues related to experimental design (e.g., dealer/player swaps/rotations).

H0: all tables generate the same revenue on average
Ha: there is at least one table for which the average revenue differs from the others

8A6

○ Since our p-value was greater than 0.05 (p = .0817), this means that the study has demonstrated there to be no association between table and revenue (in other words, the study has demonstrated, by virtue of the size of the p-value, that there is no evidence of a "table" effect where at least one table is different from the others).

○ A p-value > 0.05 (in this case, p = .0817) implies that the null hypothesis (i.e. all tables generate the same revenue on average) is incompatible with the observed data (table means of $8486.60, $8464.20, $9509.40, and $7911.40); however, a conclusion of "no effect" (in other words, an effect size of zero) cannot be determined from this result.

8A8

○ The size of the effect is not determined by the significance/insignificance of the p-value; thus, a lack of statistical significance (p = .0817 > .05) does not necessarily imply that a small effect size has been detected.

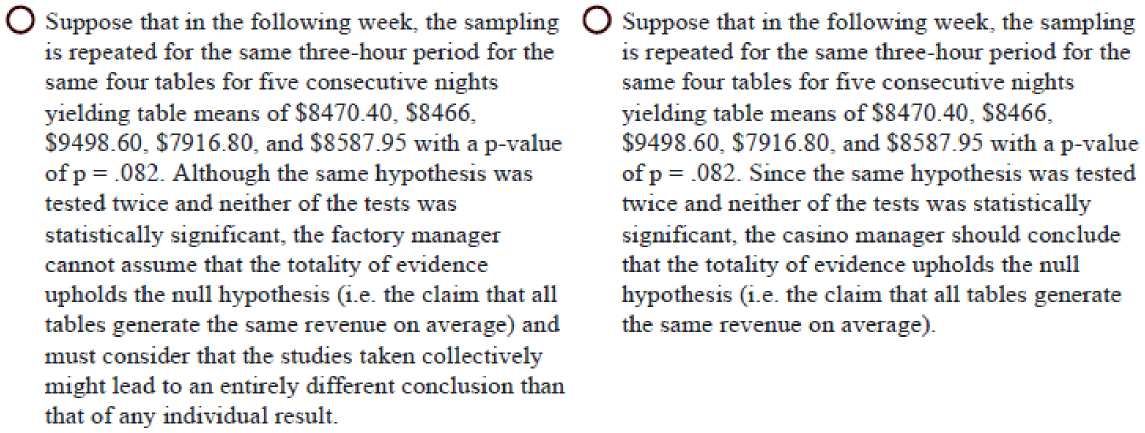
○ Lack of statistical significance (i.e. p = .0817 > 0.05) indicates that the effect size (effect of shift on call volume) is small.

8A12

○ This p-value is properly reported as an inequality: .05 < p < .10.

○ This p-value is properly reported as an exact value: p = .0817.

8A15



○ Suppose that in the following week, the sampling is repeated for the same three-hour period for the same four tables for five consecutive nights yielding table means of $8470.40, $8466, $9498.60, $7916.80, and $8587.95 with a p-value of p = .082. Although the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager cannot assume that the totality of evidence upholds the null hypothesis (i.e. the claim that all tables generate the same revenue on average) and must consider that the studies taken collectively might lead to an entirely different conclusion than that of any individual result.

○ Suppose that in the following week, the sampling is repeated for the same three-hour period for the same four tables for five consecutive nights yielding table means of $8470.40, $8466, $9498.60, $7916.80, and $8587.95 with a p-value of p = .082. Since the same hypothesis was tested twice and neither of the tests was statistically significant, the casino manager should conclude that the totality of evidence upholds the null hypothesis (i.e. the claim that all tables generate the same revenue on average).

Block 10

Scenario 8B (ANOVA): A group of paramedics does not believe that the mean number of calls received in one shift is the same for the morning, afternoon, and evening shifts. To test this claim, they record the number of calls received during each shift for seven days, resulting in the following call/shift averages: 2.57, 3.71, and 4.28 for the morning, afternoon, and evening shifts, respectively. The p-value for the appropriate test procedure was p = .039.

H0: all shifts receive the same number of calls on average
Ha: there is at least one shift for which the average number of calls differs from the others

8B2

○ The probability that chance produced the observed association (i.e. morning calls < afternoon calls < evening calls) is 3.9%.

○ The probability that chance produced the observed association between shift and call volume cannot be determined from the p-value.

8B13

○ Statistical significance is a property attached to the result of a statistical test and not a property of the phenomenon (i.e. the relationship between time of day and call volume) under investigation.

○ Statistical significance is a property of the effect time of day has on number of calls received and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.

8B3

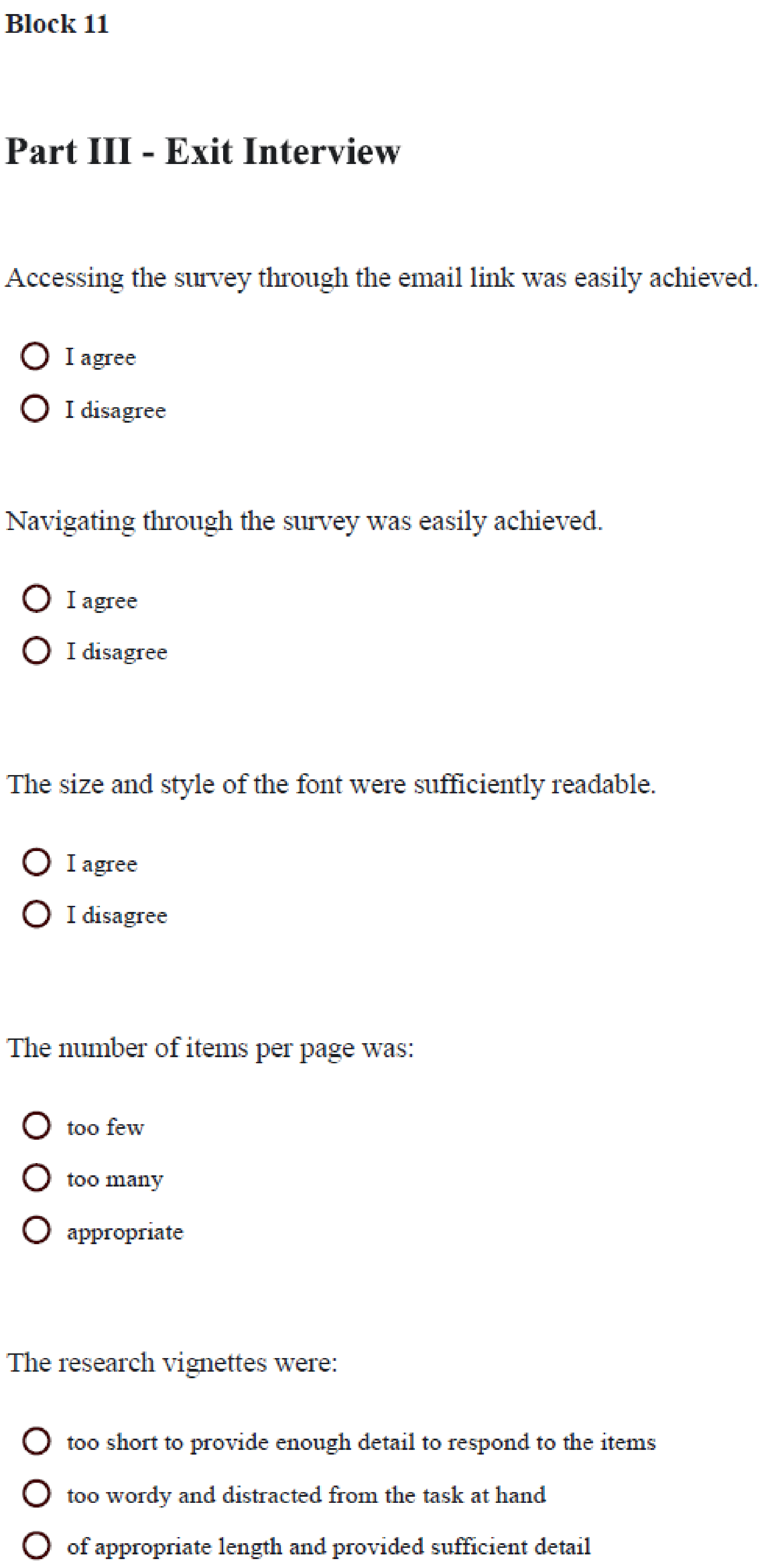
○ A significant results here (p = .039 < 0.05) means that the null is false (i.e. that all shifts do NOT receive the same number of calls on average).

○ The falsity of the null hypothesis (i.e. all shifts receive the same number of calls on average) is not established with a significant result (i.e. p = .039 < 0.05).

8B7

○ Statistical significance (p = .039 < 0.05) indicates that the observed data are unusual in relation to the null hypothesis (i.e. all shifts receive the same number of calls on average); however, the way that the data are unusual might be of no clinical interest or practical importance.

○ Statistical significance (p = .039 < 0.05) means that a substantively important relation between shift and call volume has been detected.



Block 11

Part III - Exit Interview

Accessing the survey through the email link was easily achieved.

○ I agree

○ I disagree

Navigating through the survey was easily achieved.

○ I agree

○ I disagree

The size and style of the font were sufficiently readable.

○ I agree

○ I disagree

The number of items per page was:

○ too few

○ too many

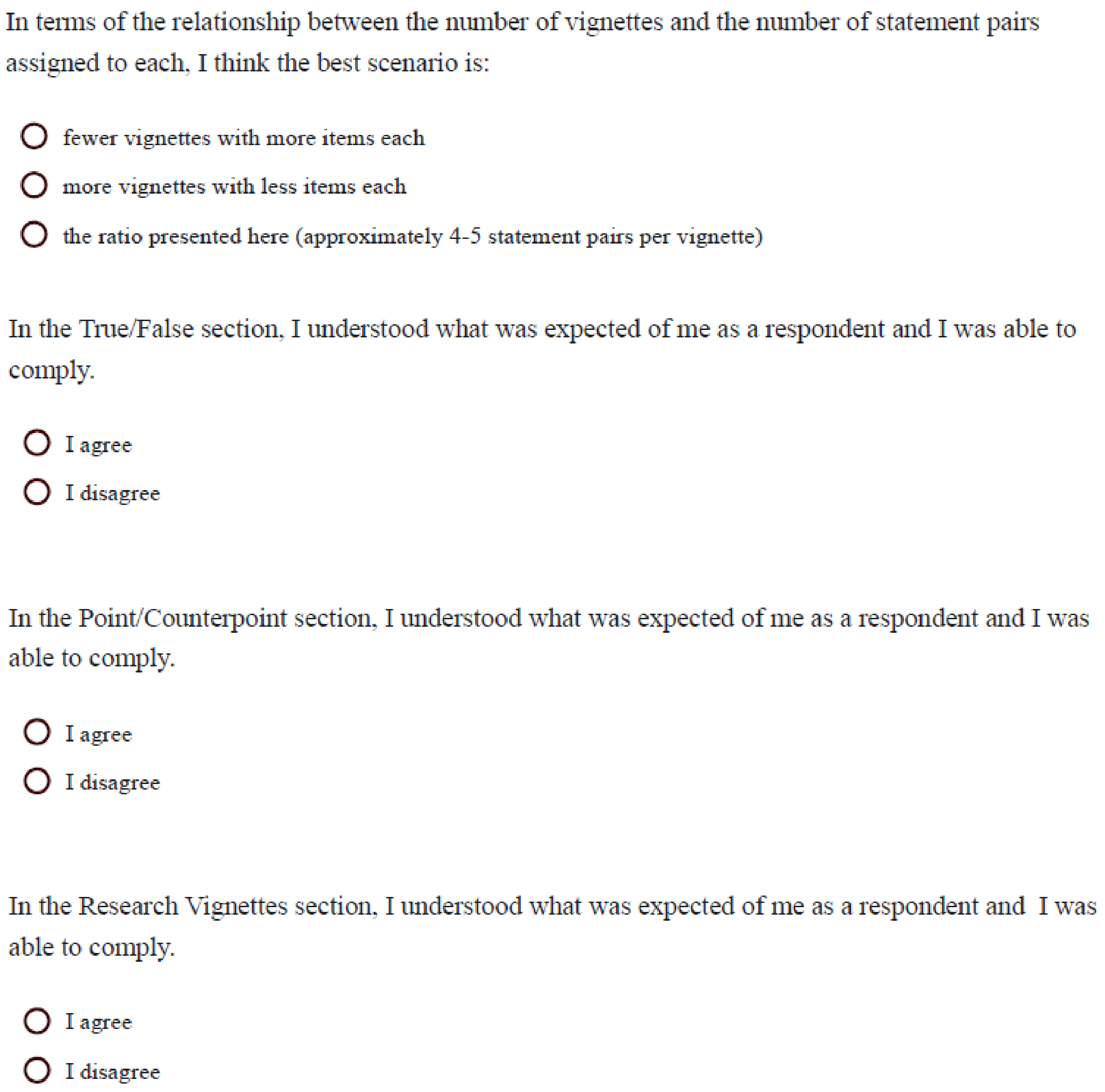
○ appropriate

The research vignettes were:

○ too short to provide enough detail to respond to the items

○ too wordy and distracted from the task at hand

○ of appropriate length and provided sufficient detail



In terms of the relationship between the number of vignettes and the number of statement pairs assigned to each, I think the best scenario is:

○ fewer vignettes with more items each

○ more vignettes with less items each

○ the ratio presented here (approximately 4-5 statement pairs per vignette)

In the True/False section, I understood what was expected of me as a respondent and I was able to comply.

○ I agree

○ I disagree

In the Point/Counterpoint section, I understood what was expected of me as a respondent and I was able to comply.

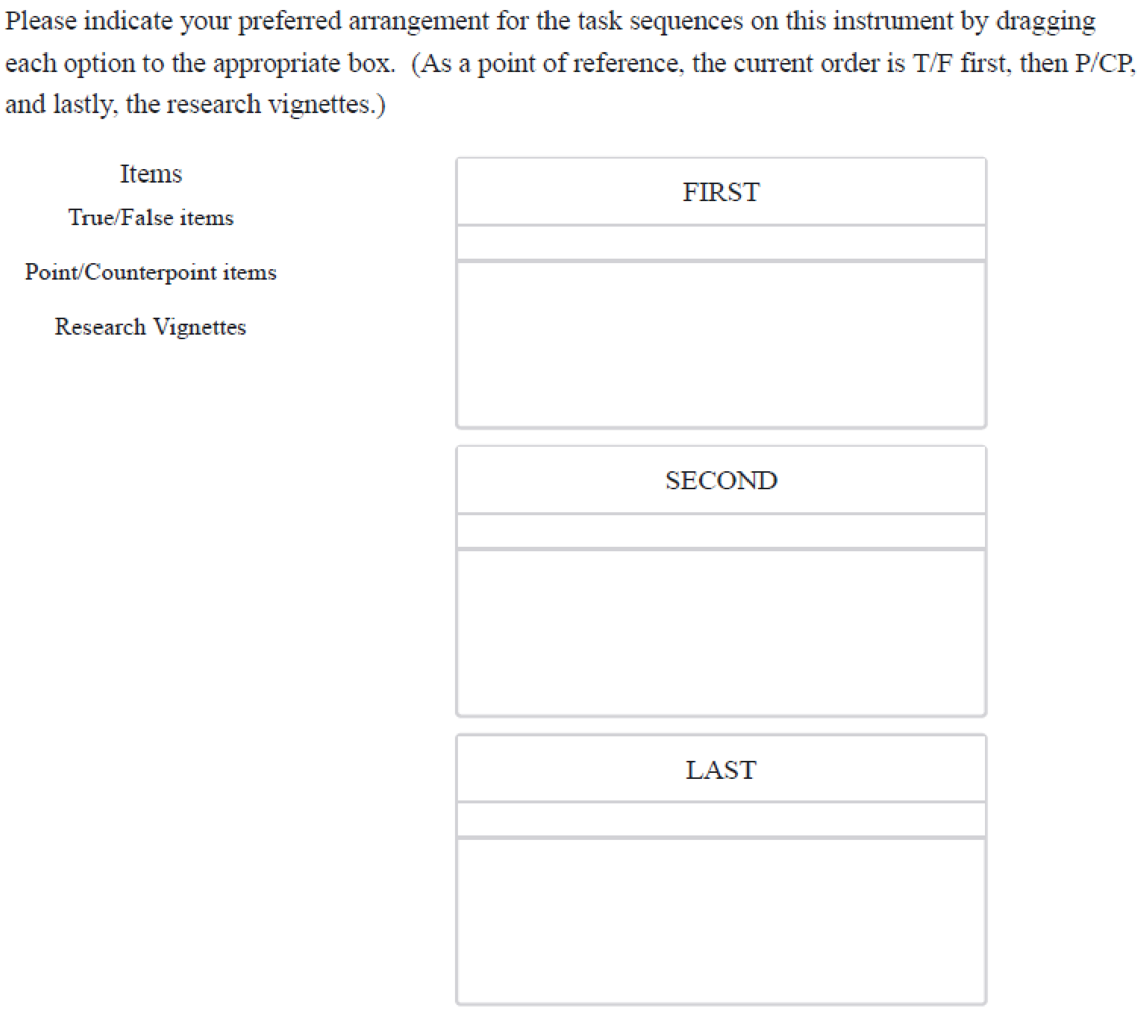
○ I agree

○ I disagree

In the Research Vignettes section, I understood what was expected of me as a respondent and I was able to comply.

○ I agree

○ I disagree



Please indicate your preferred arrangement for the task sequences on this instrument by dragging each option to the appropriate box. (As a point of reference, the current order is T/F first, then P/CP, and lastly, the research vignettes.)

Items

True/False items

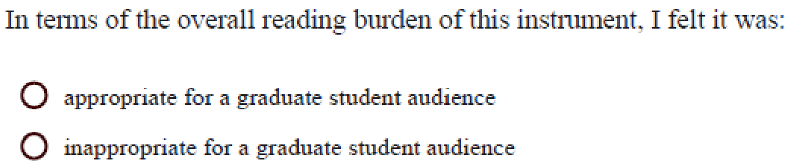
Point/Counterpoint items

Research Vignettes

FIRST

SECOND

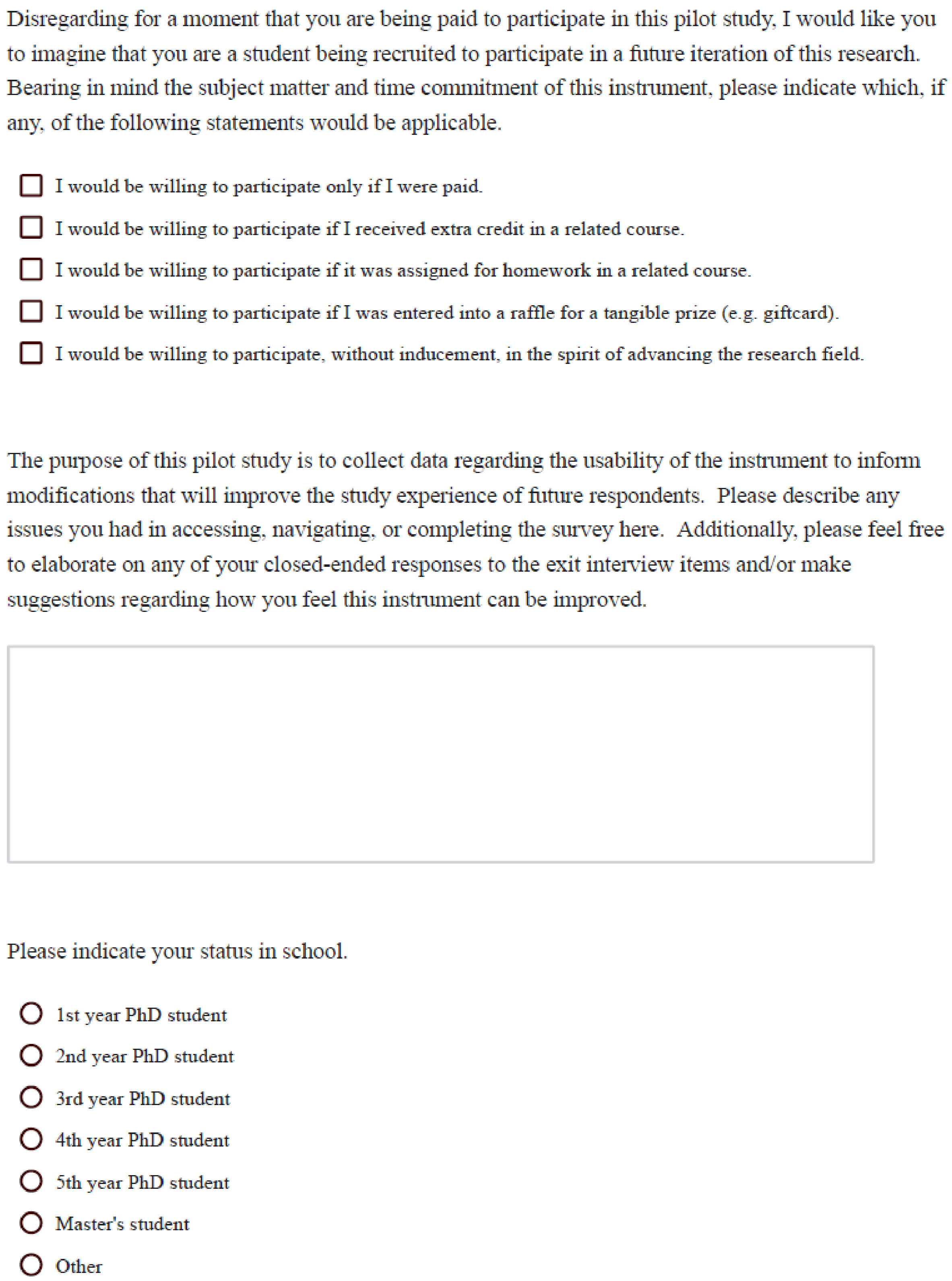
LAST



In terms of the overall reading burden of this instrument, I felt it was:

○ appropriate for a graduate student audience

○ inappropriate for a graduate student audience



Disregarding for a moment that you are being paid to participate in this pilot study, I would like you to imagine that you are a student being recruited to participate in a future iteration of this research. Bearing in mind the subject matter and time commitment of this instrument, please indicate which, if any, of the following statements would be applicable.

□ I would be willing to participate only if I were paid.

□ I would be willing to participate if I received extra credit in a related course.

□ I would be willing to participate if it was assigned for homework in a related course.

□ I would be willing to participate if I was entered into a raffle for a tangible prize (e.g. giftcard).

□ I would be willing to participate, without inducement, in the spirit of advancing the research field.

The purpose of this pilot study is to collect data regarding the usability of the instrument to inform modifications that will improve the study experience of future respondents. Please describe any issues you had in accessing, navigating, or completing the survey here. Additionally, please feel free to elaborate on any of your closed-ended responses to the exit interview items and/or make suggestions regarding how you feel this instrument can be improved.

Please indicate your status in school.

○ 1st year PhD student

○ 2nd year PhD student

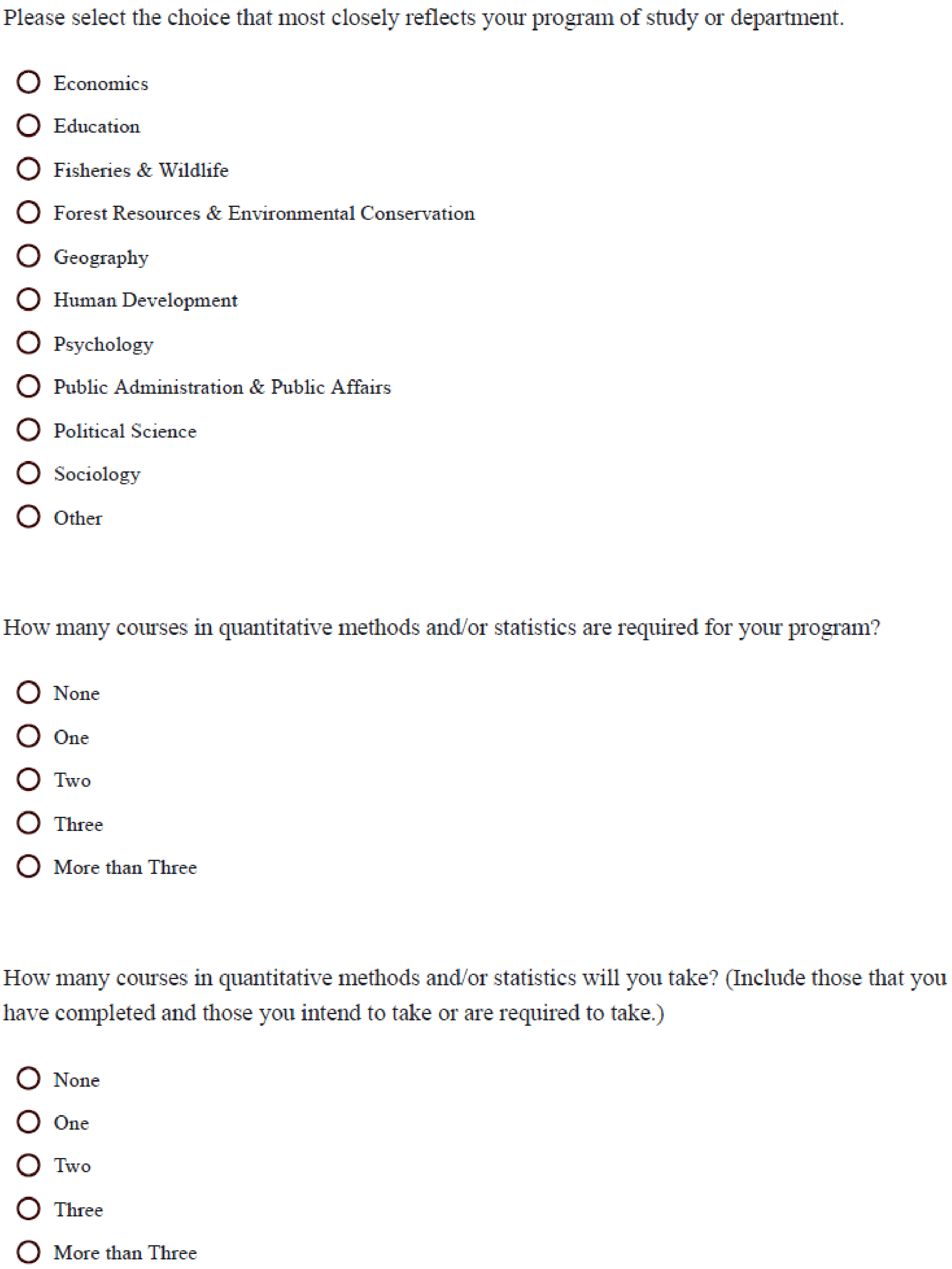
○ 3rd year PhD student

○ 4th year PhD student

○ 5th year PhD student

○ Master's student

○ Other



Please select the choice that most closely reflects your program of study or department.

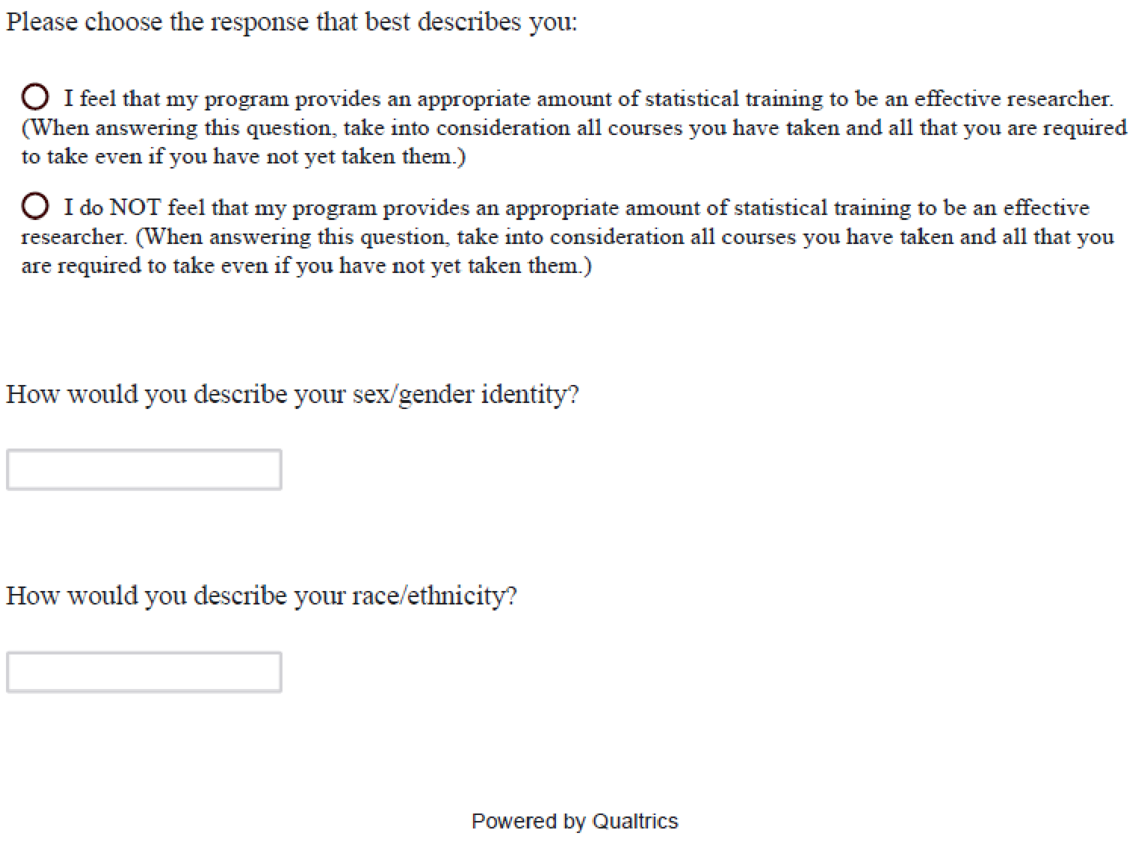
○ Economics
○ Education
○ Fisheries & Wildlife
○ Forest Resources & Environmental Conservation
○ Geography
○ Human Development
○ Psychology
○ Public Administration & Public Affairs
○ Political Science
○ Sociology
○ Other

How many courses in quantitative methods and/or statistics are required for your program?

○ None
○ One
○ Two
○ Three
○ More than Three

How many courses in quantitative methods and/or statistics will you take? (Include those that you have completed and those you intend to take or are required to take.)

○ None
○ One
○ Two
○ Three
○ More than Three



Please choose the response that best describes you:

○ I feel that my program provides an appropriate amount of statistical training to be an effective researcher. (When answering this question, take into consideration all courses you have taken and all that you are required to take even if you have not yet taken them.)

○ I do NOT feel that my program provides an appropriate amount of statistical training to be an effective researcher. (When answering this question, take into consideration all courses you have taken and all that you are required to take even if you have not yet taken them.)

How would you describe your sex/gender identity?

How would you describe your race/ethnicity?

Powered by Qualtrics

Appendix M – Phase IIA Reliability Data

Case Processing Summary – T/F Section			
		N	%
Cases	Valid	13	100.0
	Excluded[a]	0	0.0
	Total	13	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics – T/F	
Cronbach's Alpha	N of Items
0.420	16

Item-Total Statistics – T/F Section				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
M16T	9.8462	4.308	0.335	0.341
M13T	9.5385	5.269	-0.060	0.438
M11T	9.6923	5.231	-0.077	0.457
M10T	10.0769	5.577	-0.236	0.507
M14T	9.4615	5.269	0.000	0.422
M1F	9.8462	5.141	-0.056	0.460
M17F	9.9231	3.410	0.830	0.153
M2F	10.0000	5.167	-0.071	0.465
M5F	10.0769	5.077	-0.028	0.452
M6F	9.6154	5.256	-0.074	0.449
M5T	9.9231	3.744	0.626	0.235
M12T	9.6923	5.064	0.006	0.437
M8T	9.5385	4.936	0.208	0.395
M17T	9.6923	5.397	-0.157	0.476
M10F	10.1538	4.641	0.192	0.388
M4F	9.8462	3.474	0.815	0.166

Case Processing Summary - P/CP Section

		N	%
Cases	Valid	13	100.0
	Excluded[a]	0	0.0
	Total	13	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics - P/CP

Cronbach's Alpha	N of Items
0.754	7

Item-Total Statistics - P/CP Section

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
M13cp	3.6923	3.231	0.557	0.705
M7p	3.7692	2.859	0.764	0.651
M8cp	3.5385	3.603	0.477	0.726
M9p	4.0000	3.500	0.352	0.752
M15cp	4.0000	3.500	0.352	0.752
M16p	3.6923	3.564	0.346	0.751
M18cp	3.6154	3.423	0.498	0.719

Case Processing Summary - RV even

		N	%
Cases	Valid	13	100.0
	Excluded[a]	0	0.0
	Total	13	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics - RV even

Cronbach's Alpha	N of Items
0.650	18

Item-Total Statistics - RV even

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
V2A1cp	10.8462	8.308	0.685	0.577
V2A4p	10.8462	9.474	0.247	0.637
V2A5cp	11.0000	10.000	0.051	0.663
V2A14cp	10.6923	10.064	0.097	0.652
V2A9p	11.0769	9.577	0.184	0.646
V2B10cp	11.1538	10.474	-0.090	0.680
V2B11p	10.9231	10.910	-0.218	0.695
V2B18cp	10.6154	9.756	0.348	0.632
V2B16p	11.0769	10.077	0.027	0.666
V2B17cp	11.0769	8.244	0.645	0.578
V8A15p	10.7692	8.859	0.530	0.603
V8A6cp	10.6923	9.564	0.315	0.631
V8A8p	10.7692	9.526	0.265	0.635
V8a12cp	10.7692	9.526	0.265	0.635
V8B2cp	10.8462	10.641	-0.139	0.684
V8B13p	11.0000	9.167	0.318	0.627
V8B3cp	11.0000	8.500	0.551	0.593
V8B7p	11.0000	8.333	0.612	0.584

Case Processing Summary - RV odd

		N	%
Cases	Valid	13	100.0
	Excluded[a]	0	0.0
	Total	13	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics - RV odd

Cronbach's Alpha	N of Items
0.773	18

Item-Total Statistics - RV odd

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
V3A1cp	10.6923	13.564	0.413	0.758
V3A17p	10.9231	13.077	0.509	0.750
V3A3cp	10.8462	14.974	0.003	0.790
V3A9p	10.6923	14.897	0.035	0.785
V3B4cp	10.7692	12.692	0.640	0.739
V3B12p	10.5385	14.603	0.179	0.773
V3B16cp	10.9231	12.910	0.557	0.746
V3B15p	10.8462	12.974	0.538	0.748
V5A10cp	10.9231	15.744	-0.184	0.803
V5A7p	10.8462	12.641	0.636	0.739
V5A11cp	10.8462	13.974	0.261	0.770
V5A18p	10.6923	12.564	0.723	0.734
V5A14cp	10.4615	14.603	0.272	0.769
V5B2cp	10.9231	12.744	0.606	0.742
V5B5p	10.8462	13.308	0.444	0.755
V5B6cp	10.6154	13.090	0.622	0.744
V5B13cp	10.5385	16.269	-0.381	0.803
V5B8p	10.6154	13.590	0.456	0.756

Item-Total Statistics - Full Instrument

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
M16T	37.1538	93.474	0.439	0.880
M13T	36.8462	98.808	-0.156	0.886
M11T	37.0000	101.167	-0.378	0.890
M10T	37.3846	100.590	-0.278	0.890
M14T	36.7692	98.026	0.000	0.884
M1F	37.1538	92.641	0.526	0.879
M17F	37.2308	91.526	0.628	0.877
M2F	37.3077	99.564	-0.175	0.888
M5F	37.3846	95.423	0.237	0.883
M6F	36.9231	94.910	0.406	0.881
M5T	37.2308	92.192	0.558	0.878
M12T	37.0000	100.167	-0.266	0.888
M8T	36.8462	95.308	0.488	0.881
M17T	37.0000	100.500	-0.303	0.889
M10F	37.4615	96.936	0.091	0.885
M4F	37.1538	90.808	0.721	0.876
M13cp	37.0769	91.744	0.658	0.877
M7p	37.1538	90.641	0.739	0.876
M8cp	36.9231	94.910	0.406	0.881
M9p	37.3846	91.590	0.638	0.877
M15cp	37.3846	93.256	0.461	0.880
M16p	37.0769	94.744	0.326	0.882
M18cp	37.0000	93.500	0.511	0.879
V2A1cp	37.0769	91.744	0.658	0.877
V2A4p	37.0769	94.910	0.308	0.882
V2A5cp	37.2308	97.359	0.039	0.886
V2A14cp	36.9231	97.910	-0.003	0.885
V2A9p	37.3077	94.897	0.283	0.882
V2B10cp	37.3846	97.256	0.051	0.885
V2B11p	37.1538	97.808	-0.004	0.886
V2B18cp	36.8462	95.641	0.425	0.881
V2B16p	37.3077	96.231	0.150	0.884
V2B17cp	37.3077	91.064	0.676	0.876
V8A15p	37.0000	93.500	0.511	0.879
V8A6cp	36.9231	96.244	0.223	0.883
V8A8p	37.0000	94.500	0.391	0.881
V8a12cp	37.0000	96.000	0.213	0.883
V8B2cp	37.0769	98.410	-0.065	0.887
V8B13p	37.2308	93.526	0.422	0.880
V8B3cp	37.2308	91.359	0.645	0.877
V8B7p	37.2308	91.526	0.628	0.877
V3A1cp	37.0769	94.910	0.308	0.882
V3A17p	37.3077	93.064	0.469	0.879
V3A3cp	37.2308	96.026	0.170	0.884
V3A9p	37.0769	97.410	0.041	0.885
V3B4cp	37.1538	90.308	0.775	0.875
V3B12p	36.9231	97.077	0.109	0.884
V3B16cp	37.3077	91.231	0.658	0.877
V3B15p	37.2308	91.692	0.610	0.877
V5A10cp	37.3077	99.897	-0.206	0.889
V5A7p	37.2308	92.026	0.576	0.878
V5A11cp	37.2308	95.359	0.237	0.883
V5A18p	37.0769	91.744	0.658	0.877
V5A14cp	36.8462	97.308	0.117	0.883
V5B2cp	37.3077	92.897	0.486	0.879
V5B5p	37.2308	92.859	0.490	0.879
V5B6cp	37.0000	92.167	0.673	0.877
V5B13cp	36.9231	100.244	-0.314	0.888
V5B8p	37.0000	91.833	0.714	0.877

Reliability Statistics

Cronbach's Alpha	N of Items
0.883	59

Case Processing Summary

		N	%
Cases	Valid	13	100.0
	Excluded[a]	0	0.0
	Total	13	100.0

a. Listwise deletion based on all variables in the

Appendix N – Video Transcript Log

T/F	Selection	Problem	Understands Idea	Transcript Excerpts (Abigail)
MI10T	incorrect	error rates	likely not	*I don't know much about error rates. Is this like sampling error types of things? [Nope…they are referring to Type I and Type II errors, have you heard of those?] Yes, I did not know that at all. (Repeats question to herself). So Type I error, say you have an alpha of .05...Type I error is the likelihood that we reject the null when we should not have rejected the null hypothesis. I mean usually we are stating a kind of threshold that we are allowing, I don't know that, I don't see what that has to do with any of this [what the question is supposed to be asking is, if I set my alpha at .05 and do a single hypothesis test does that mean I have a 5% chance of committing a Type I error on that test] No [or if I do 100 tests, I would not have committed a Type I error on 95 of them and I would have on 5 of them, but the tricky part is that I don't know which 5 are wrong. The whole point of the question is whether it is a single probability or a long run frequency]. It is not single, or is it? I never thought about it this way before, either way. So, if there is - I feel like it should be over the long run. Basically, the definition...if we get .05 for our p-value, we talk about that being the likelihood that we get this extreme of a result by chance, so I suppose if you have 3%, this is the likelihood that they are actually the same and this really extreme result happened by chance. So that would speak to more like this particular instance. So, I guess I disagree? I never thought about it this way at all. [Do you understand what the question is asking after I clarified it? Or you understood it originally?] I needed the 'Type I/Type II' clarification, now I understand the question but am not sure which one is correct.*
MI10F	Incorrect		yes	*I believe that is true because, yeah, Type I error is supposed to be about the alpha value, so that should be true*
MI4F				*No, I disagree*
MI17F	Correct		maybe? Selection is based on decision being the same and not on p being the same, but thinks false for the wrong reason	*I don't think the p-values have to be the same, [they could be confirmatory even if the p-values are not the same?]…well, 2 cases still is not very confirmatory to me…I suppose there is a difference between confirmatory and truth, so I was speaking in terms of truth and confirmatory is different. If you have 2 papers having similar outputs, similar conclusions then that would be confirmatory but I still don't think the p-value is the point. The fact that you are rejecting the null is the point. I feel like I disagree because it is not about having the same p-value, it is about your conclusion.*

T/F	Selection	Problem	Understands Idea	Transcript Excerpts (Abigail)
MI16T	Correct		emergent understanding: knowledge that seemingly contradictory results are not necessarily contradictory, but cannot fully articulate this idea	*We are talking about the same effect with these observed associations but one is fail to reject and one is reject? [Exactly] well yeah, I suppose it would depend on your sample size…I don't know enough about effects. I know they are a thing but I don't know much other than that. I would assume it is possible for that to happen.*
MI13T	Correct			*Yes, statistical significance is going to be related to the test which has to do with sample size and has nothing to do with the underlying relation so I agree.*
MI11T	Incorrect		likely not	*Because you are just saying that, you are referring to a specific p that is referring to a cut-off point, it doesn't have to be observed necessarily, so I would disagree.*
MI14T	Correct		possibly not, doesn't seem to know difference between 1 and 2-sided p or why both would exist…comments corroborate that she has never seen/done this before	*yeah, I mean if you are talking about if something is equal/not, then you need a 2-sided test, but if you have very good evidence that it is >,< you should use a one-sided test…[does the p have to match they hypothesis?] I don't know what that means…still not sure what that means, if you are using a 1-sided test, your 5% rejection region is going to allow you to have a much closer to the middle critical value, so it will change what your critical value is...[do you know how to calculate a 2-tailed p-value?] I know how to calculate a critical value. [more clarification] I know there is a formula for calculating p-values, but I don't have it memorized, it is related to like you have your mean, depending on direction you have your +/-, then you have z, t and it is related to the # of standard deviations and it should be st dev over root n...I still don't understand...so when you are calculating your p-value, you are going to be choosing which z or t value according to if it is 1-sided or 2-sided, is that what you mean?*
MI1F	Correct			*No, (pause), No.*
MI12T	Correct		yes	*That is probably healthier, if you get 10^-13 nobody cares…but if it is reasonable decimals, it is better to give exact values for comparing and transparency in reporting.*
MI8T	Correct		yes	*I suppose even large effects could be masked by noise in the case of really small samples, but especially small effects could be masked, but I guess I still agree*
MI5T	Incorrect	look at wording of negation?		*Oh, accepting the hypothesis, so we are talking about accepting the null hypothesis. No, you cannot assume that. So if I cannot assume that, the conclusion cannot be deduced, therefore I agree with the statement. [Were there too many 'not's' in there?] That was a little bit hard to parse.*

T/F	Selection	Problem	Understands Idea	Transcript Excerpts (Abigail)
MI2F	Correct		yes	*p-value is about, if I have .03 as our p, that means there is a 3% chance that the means, if we are talking about means…the probability that we observed this outcome, whatever it was, and the means are actually the same…and if it is a really extreme value, having a small probability means that it is unlikely that those are the same...probability that chance produced the observed association - I don't feel that is the same thing, so I disagree*
MI6F	correct	demonstrated	yes	*No, they have not proved there to be no association…well, demonstrated, bother…[I don't want you to second-guess, you picked the correct answer. Do I need to change the word 'demonstrated'?] well, demonstrated seems like evidence of, so…in my head I was thinking demonstrated versus proved [I think I was using demonstrated as a synonym for proved]. So, I originally read it as proved and was going to choose and then I was backing off as well what does demonstrated mean? Those aren't quite the same in my head.*
MI17T	Correct			*not sure what 'observed association' is [gives example to clarify] I fully recognize my stats knowledge is not the strongest, so if I didn't know, I would start asking people…I feel like I don't focus on what the p-value is exactly, it is the reject/fail to reject more...have basically the same conclusion from these two studies that there is a difference but the size of the difference is different...well, yeah, I would think so because there is always some amount of error as you are doing stats so it would be possible for 2 studies to get different differences even just for sampling reasons, so I guess I agree*
MI5F	Correct			*so, we are never allowed to accept the null hypothesis, that is just not good...in terms of if we are getting a giant p-value, then it is possible that we are thinking about this wrong and that there is actually more validity or it is actually true or it is possible that the null hypothesis is true - we could not conclude that it is definitely true but I guess it would be evidence in favor of thinking about it but you still could not come to a firm conclusion so I guess I would agree...[I think what we are trying to do with that item is say, a large p is giving us evidence to support the specified null as opposed to a lack of evidence to support the alternative and technically there are many null hypotheses that we could have chosen that may or may not be more appropriate than the one we chose so getting a large p...I think the words 'in favor of' is causing trouble because I think you interpreted it differently than maybe I wanted people to]...on a side note, based on my Quant I class, I would definitely say that I agree but I don't know if I buy what they said in my Quant I class. At some points in that class, they were talking about accepting the null, which from what I heard having taken other stats classes, is the immediate, 'no, you should never say that'*

P/CP	Selection	Problem	Understands Idea	Transcript Excerpts (Abigail)
MI3cp	correct		yes	*So, this is going to the Type I error, I think, right? You end up concluding whether or not it is actually true. A significant test result does not necessarily mean that the null is false even though that is going to be our working assumption moving forward...so I would say the second one.*
MI7p	correct			*Not necessarily (upon reading second one) - first one.*
MI8cp	correct	proper negation?	yes	*Oh, 'has been detected'...[does that trip you up?] That doesn't seem like version and negation. I was expecting it to conclude 'does not necessarily imply that the effect size is small'. [assume it is a correct negation] There are enough other factors going on that that is not enough to assume that, so I choose the second one.*
MI9p	incorrect	should be 'than that' not 'that that'	possibly, here she seems sure that the second one is correct (evidence of a misconception), but previously in this interview, she defined the p-value as the likelihood of 'this extreme of a result' indicating that she does understand this distinction. I think this is a limitation of her knowledge in that she can recite the definition but not necessarily apply it in context?	*The second one makes more sense to me. By 'observing the data we observed?' [clarification: the chance of getting the data you got if the null were true] Yeah, yeah.*
MI15cp	incorrect			*uphold seems too strong, for one thing. Because upholds sounds 'provish' to me [and it is supposed to]. This does not prove the null hypothesis, it may lend credence to testing what you were testing, but you have not proved the null hypothesis. I guess that means the first one...[clarification regarding combining evidence from multiple studies] I wouldn't say it is a done deal, but I would say that the null is probably true at this point. [Your problem with this question is that you know we cannot accept the null and you think the question is saying that we should accept the null - that is not exactly what we were going for...]*
MI18cp	correct	incorrect is spelled wrong		*just because you observe a small p-value doesn't mean the next study will, it depends on a whole host of factors, you can't assume automatically you are going to get a small p-value again*

P/CP	Selection	Problem	Understands Idea	Transcript Excerpts (Abigail)
MI16p	incorrect	clarify what is meant by 'contradictory'	Not sure, her initial reasoning would suggest choice B but then her later clarification would suggest choice A. When saying that, she would pick the more trustworthy study, she is indicating the results are in conflict, but when suggesting that future research is necessary, that seems to indicate that they are not necessarily contradictory. Your first response of I would look at both and pick the one that is more trustworthy leads me to think you would choose B - that they are contradictory, right? So you want to pick the one you agree with. But, if you couldn't decide between the two as being more/less trustworthy, then you would just do another study which leads me to believe that you would choose A because you are saying that they are not necessarily contradictory. She has made a case for both answers. Limitation of her knowledge or of the item itself?	*I want to say true, but I should read the other one. (reads to herself) Not necessarily no. Because there could be underlying issues, there could be explanations for why you would get both sides. Well, I guess that is kind of contradictory. [The decision is contradictory - reject/not reject, but does that mean] That doesn't mean the validity of the study is called into question. (clarifies difference between decision and conclusion) I think that is what I was originally thinking about, if you have p-values on opposite sides of .05, there are a host of things that could be leading into that being the case from sample size to sampling error or who you happen to sample, so it is perfectly possible for that to happen. In terms of your decision moving forward, at that point you start looking at those ideas of sample size and methods type things and using that to decide which study seems more valid or likely, and based on that you make a decision in terms of what I believe to actually be the phenomenon going on...but it is possible, even if you did everything right, maybe this study had a small sample size...[don't assume that either study did anything invalid, both p-values are legitimately obtained]. I guess to some extent they at least provide evidence and counter evidence so I would go with the one that seems to be more likely or that seemed to have the larger n or the better looking questions. I would start using those to make a decision about which one I believe and if I felt they were pretty even, I would say further study is needed...As a researcher, I would need more information than just the p-value to make a definitive conclusion.*

RV2	Selection	Problem	Understands Idea	Transcript Excerpts (Abigail)
V2A1cp	Correct		likely as she correctly identifies that the 'point' wording was really getting at the truth of the null	*(reads the 'point' option) i.e. that the null is true, that probability cannot be determined from the p-value because the p-value is referring to the probability or the chance of them actually being - No, that cannot be determined.*
V2A4p	Correct			*I just said that's not true so the truth of the faculty's claim is not established.*
V2A5cp	Correct			*I am not sure what the end of that sentence means. I am going to ignore the end of that sentence since the first one made sense. Because this test produced a large p-value, what is large exactly? [large is subjective but in this case, I am saying that p is much larger than .05] This is a random sample of 12 professors, so there is a decent chance that there is a difference (i.e. less than 15), but that is not enough to support the professors claim yet...we did get 14.8 which is less than 15 but it is not big enough to be significant for our purposes...so the second one.*
V2A14cp	Correct			*(reads the 'point' option) Why? We had a discussion about this earlier. (Rereads scenario) So like negative? Is that what you mean? If we are looking at least or less than, that should be 1-tailed [So, should the p-value be 1-tailed or 2?] 1-tailed. So it should be reported as-is. I assume that's what that means.*
V2A9p	Incorrect	identifies typo (*is 28.8%* is missing from the stem)	same issue as MI9, now twice she makes the same mistake, so likely that she does not grasp the idea of p being a tail probability and not just the chance of the observed results alone.	*[clarifies the wording] My problem is that there is no verb. (re-reads the stem) Yes.*
V2B18cp	Correct			*No, we just said the previous example…that is a perfectly reasonable thing to happen.*

RV2	Selection	Problem	Understands Idea	Transcript Excerpts (Abigail)
V2B10cp	Incorrect	emphasize 'long run frequency'? "skipping over the parentheses is a bit hard when it is this long"	"Not sure I completely understand that concept and that is part of the problem"	*(reads the 'point') Pretty sure that is true. (reads 'counterpoint') I am assuming from this that our cutoff was .05 so we are allowing ourselves a 5% chance of committing a Type I error going in so that leads me to say the first one. [The question is not asking if alpha = .05, is there a 5% chance of a Type I error because that is true by definition, that is what alpha tells us. Does alpha tell us that on the next test there is a 5% chance of making that mistake or in the long run looking at all the tests collectively, 5% of them will be Type I errors?] You wouldn't necessarily have 5% that would be Type I errors because there are other types of error...[clarifies again, Type I error "in this instance" versus "long run frequency"] Yeah, for any particular instance, you don't know if you committed an error, so I guess it would be a long term thing.*
V2B11p	Incorrect	same meaning = same conclusion? Same decision?		*I mean, .011 gives you more information but you would come to the same conclusion, so I guess the second one. [If one gives you more information, can they convey the same meaning?] I guess when I'm reading the same meaning, I am thinking the same conclusion [you are thinking reject/not reject decision, but that is a decision not a conclusion] Correct. .011 is more information.*
V2B16p	Correct			*Considering you had a smaller sample size, I wouldn't be shocked that you were no longer significant even though you ended up with a similar mean and standard deviation…roughly the same output and I would not be surprised if now we have a larger p-value because we have a smaller n, in some senses, I could interpret that as confirmatory, especially if we are allowed to pool things, I would be inclined to pool things and say see look at the whole thing, so first one.*
V2B17cp	Incorrect		unclear…she is correct in assuming that 'results' can mean many things, but she is focusing attention on the decision (reject/not) and not the conclusion…with a much larger sample, the reduced effect should be interpreted as contradictory to the previous result (I think the item is correctly capturing her misconception)	*(reads 'point') That is a little sketchy (reads 'counterpoint'). The word results is going to determine what I say here. If results means conclusion, then I guess it is like what conclusion are we looking at? Yes, we still have a waittime of over 30 minutes and it is still significant so there is a decent chance that in fact we still have a waititme that is bigger than 30 minutes, and that is all fine but the size of the waittime is a good deal smaller here, or at least was a few minutes smaller...the fact that they both had nearly, I mean the number for the p-value isn't really what's important, but the idea that there is a decent chance that you wait for more than 30 minutes is being confirmed, the idea of the size of that difference is not being confirmed. We have a good amount of variability there so if I am looking overall at this idea of is there a decent chance that people wait more than 30 minutes, then I consider that confirmatory. If I am looking at the size of the effect, this is not confirmatory. [look at it strictly based on the framing of the hypothesis] Then, I guess it is confirmatory.*

RV8	Selection	Problem	Understands Idea	Transcript Excerpts (Abigail)
V8A6cp	correct			*(reads 'point') That seems a bit bold. (reads 'counterpoint') That one.*
V8A8p	correct			*We have what 4 tables? So, the number of tables you have or the more piles you put something in, will dilute the power of your test. That might not actually be relevant. [clarifies by saying is the size of the p-value directly related to the size of the effect] No. If our p value is greater than .05, then we haven't even detected an effect size, so just based on that portion of the statement, I should choose that it does not imply a small effect size has been detected.*
V8A12cp	correct			*I think it makes more sense to say the exact value since you are giving the person more information.*
V8A15p	no response	looks long on phone, issues with scrolling; grand mean / table means label missing?	unable to determine, error in question presentation prevented respondent from answering	*I assume collectively means Table 1 all night, Table 2 all night…No, especially if I actually do look at the values - I have to scroll that is slightly annoying - (discussion of phone vs computer presentation)…the same 4 tables, there are 5 numbers here, is one of these supposed to be the grand mean? So this is probably ad hoc stuff getting in the way, is this the first one supposed to be the grand mean or the last one?*
V8B2cp	incorrect	Ordering was confusing.		*My understanding of ANOVA is not as strong as some things. So the alternative hypothesis is just that there is a difference. This is putting them in an order. [I am only ordering them because the means are not equal, I am not implying that the test ordered them...we are not assuming a post-hoc result here] The p-value is telling us the likelihood of getting the result we had and the null being true...I guess I should say that's true then.*
V8B13p	correct			*(reads point) Yeah. (reads counterpoint) That statement doesn't make any sense so I assume it is wrong. My immediate reaction to the first one is I believe this to be true, this makes less sense to me and I don't think it is true. [Does the significance go with the phenomenon or the test?] Test.*
V8B3cp	correct	typo (need result not results)		*This is where established…[think proved]…I mean you do one statistical test, maybe you had a weird sample, there is a whole host of factors leading into this, so we have not proved definitively that this is the case, so second option.*
V8B7p	correct			*So, what does clinical interest or practical importance mean? [clarification provided] I would have to be a paramedic to know that, I suppose a difference of one call per shift would not necessarily change how many people you are going to hire, that is probably similar enough to not change anything...I guess I would say the first one, it is statistically significant, I don't know that it has clinical interest, it might, but I don't know.*

T/F	Selection	Problem	Understands Idea	Transcript Excerpts (Babs)
MI16T	correct	perfect agreement	yes	*not sure what you mean by 'perfect agreement'…are you telling me that the actual data points are identical? Or are you telling me that the studies are done on similar populations? That question is confusing me. If I stop and think, I think I can answer it. Yes, because it depends on the variation within each data set. I think that 'agreement' is a very general term and it could mean many things and you have given me one example of what that means but you haven't said that's the only thing.*
MI13T	correct			*it is really just saying, here is my distribution within the population…how much of those possible outcomes am I willing to say could fall within there basically…[where is significance ascribed to?]…I am saying to the test itself, but then we make the jump to the other thing (i.e. phenomenon) because without that, why would we even do the statistics? But I think that is the trick...*
MI11T	correct		maybe? P = .05 and p = .02 are not the same thing, "yes, but they are going to come to the same conclusion"	*Ok? I think I know what that means. Ok, I don't know the exact value of p from this statement, so just from that, I know that p could be any value < .05, not that p is actually many values, it is just that I not reporting the exact value…which is just a mathematical statement...I can do away with all the rest of this sentence because these are just mathematical statements that don't mean the same thing at all.*
MI10T	incorrect	long-run frequency	maybe? She can articulate the same concept with CI, but doesn't realize it applies here as well.	*"this may just be because I don't know this" [explains the term] "I don't understand the difference between the two. That is what those words mean to me, but I don't understand…probability always makes my stomach flip…all those probability things, I don't quite get them. I definitely think it is the knowledge, because that is definitely what I thought you meant by those words.*
MI14T	correct	hypothesis of practical interest	yes	*I think I know what that means, but I wonder about that statement…as opposed to the null? Is that why are you differentiating? That is what I thought it meant, but it is confusing*
MI1F	correct			

T/F	Selection	Problem	Understands Idea	Transcript Excerpts (Babs)
MI17F	incorrect			*see, this is another one…'the researcher should interpret these results as confirmatory'. I think what you mean by 'confirmatory' is certainty, or it could be that these agree…[after an explanation], the data from the studies are telling you the same thing? I probably don't know this...so the same hypothesis is tested twice, so you are saying it is tested in the same way? they have different data sets, but they were arrived at in the same way and so, ok...[if the p-values are identical, is it confirmatory or is there room for the fact that it is possible that your finding is conveying different info or is conflicting?] I replicate the study...it doesn't really matter that the p-values are identical, what matters is that we had the same cut-off value and that both studies fell above/below that line...so I think it is a limit of my knowledge*
MI2F	incorrect			*that one makes sense*
MI5F	incorrect			
MI6F	correct			
MI5T	correct	definitive conclusion	yes	*so, by definitive conclusion, are you saying that you are certain that your hypothesis is true or are you accepting that hypothesis? [sometimes the decision and conclusion mean the same thing, sometimes they don't] I think my trouble is…by definitive, you can be certain in your conclusion, like absolute truth? [If the p were very tiny, would that prove the null were false?] No, absolutely not...So, your wording is a 'definitive conclusion', does 'definitive conclusion' mean I have absolutely no doubt about what I am saying, absolute truth...No matter how large the p-value, one cannot be certain about one's conclusion (for instance, accepting a particular hypothesis)*
MI12T	correct		yes	*so, if they are like .0003 but my precision is only to the thousandth place…yeah, I guess I agree. [If you get .03 do you report that or an inequality?] Depends on what I am reporting…if I give the test, I give the exact value, but in my table, I just star those < .05*
MI8T	correct			*that one makes sense*
MI17T	correct	observed association	yes	*meaning, like, correlations? Let me make sure those words mean that…observed association is a nebulous term to me…[offer test statistic instead]…yes, that would make more sense and I might give a few examples*
MI10F	correct		misconception	*ok, I think that is true [re-reads]…No, the chance is whatever 'p' is, it is not just .05*
MI4F	correct			

P/CP	Selection	Problem	Understands Idea	Transcript Excerpts (Babs)
MI3cp	correct			
MI7p	correct	clinical interest		*2 things you might mean: just because I get a sig. p does not necessarily mean that what is going on with my data is important; the other thing I could think is that I asked the wrong question*
MI8cp	correct		yes	*I know with a significant p, you could still have a small effect size*
MI9p	incorrect			*found typo*
MI15cp	correct	totality of evidence / upholds the null	maybe? [seems to be hung up on pooling data vs. considering aggregate conclusion based on a handful of studies (not pooled)]; seems to understand how the file drawer effect could be leading to abandoning investigations when they shouldn't [maybe think of like a scale balance, which direction does scale tip with the new evidence?]	*yeah, I think this is the same thing I had a question about before….[doesn't seem to understand 'totality of evidence'] …oh, so literally if I pooled ALL the data and tested again…each study might not be sig on its own, but if you pooled it, because of the variance, etc. Wihtout you explaining that, I would not have figured that out. It could be either one (limit of knowledge or bad wording). When you say totality of evidence, as in I amgojng to combine all data and test again, you might get a different result - that is easy for me to see. But when you say, and , maybe this is it, upholds the null hypothesis, it is worded the same, but you are applying it to different groups...maybe I am confusing null hypothesis and conclusion...Really maybe what you are saying is how much can you generalize? You can never say the null is true (even upon repeated sampling). The more you see it, the more likely it is, but it doesn't ever mean that it is absolutely that way.*
MI16p	correct			*I think I know what that means, p is not a magic number*
MI18cp	correct			*that one is really, really clear*

RV3	Selection	Problem	Understands Idea	Transcript Excerpts (Babs)
V3A1cp	correct		maybe?	*I think this is that same thing that I am not sure about. [rephrases] Then I would choose that it does, but I am not certain. I think this is not a question thing, but a lack of knowledge.*
V3A17p	incorrect			*Not that I am certain, just that they agree with each other. Your question is trying to get at that magical p-value, not that I am pretty certain that if I did this a bunch of times, I would get this same result (not the same mean), but I am saying that the population mean is 2.0 cm. [they have the same p but are they conveying the same information?] I would say that they are conveying the same, but not because it is .048, but because I am pretty sure the bolts are within the tolerance...because the mean value is close to 2.0...because my null is that the mean = 2.0 cm and when I get my p-value, it is small, so that means there is a small chance that if I tested over and over again, this would not come out the same way. Most of the time it is going to come out the same way , so I would make the same conclusion in that my machines are working just fine. [means are on opposite sides of null value, but p is the same; are they conveying the same info? are they in agreement]. I would say they are in agreement because they are within tolerance...*
V3A3cp	correct			*I think you are getting at certainty here, can you be certain because you have that magic low number?*
V3A9p	incorrect			*I understand what these mean and I am not sure so I am going to pick an answer.*
V3B4cp	correct			
V3B12p	correct			*oh, you are talking about the original p-value. Since I just read about those other studies, it might be nice to know this was the original.*
V3B16cp	correct		yes	*This makes sense, maybe it makes sense. I think you are getting at that magic number p-value.*
V3B15p	correct			*This is the same thing we have been getting at so it makes as much sense as it makes.*

RV5	Selection	Problem	Understands Idea	Transcript Excerpts (Babs)
V5A10cp	correct			*I am going to pick that one because we discussed that.*
V5A7p	incorrect			*I think those are clear.*
V5A11cp	correct			*Ok, I think I know what you mean. I think you mean exact same meaning not that you don't come to the same decision. You are using 'decision' and 'meaning' pretty specifically but that is not clear to me.*
V5A18p	correct			*I think I know what you are getting at there.*
V5A14cp	correct			*I think it is that one? No, I understand…it is that same thing 'of practical interest'…I would say to maintain compatibility with the null hypothesis*
V5B2cp	correct			
V5B5p	correct			*One way you state score improvements is with < and one way you write it out in words, it would be nice if these were the same. I would go with the words because that is how you stated it in the hypotheses.*
V5B6cp	correct			
V5B13cp	correct			*Yeah, that one is clear.*
V5B8p	correct			*This makes sense. People will know what effect size is.*

T/F	Selection	Problem	Understands Idea	Transcript Excerpts (Dana)
MI16T	correct	perfect agreement?	maybe?	*oh, identical observed associations - got it; now that I understand, I agree*
MI13T	correct			
MI11T	incorrect	long question;	no	*are we trying to determine if p-values and p-value inequalities are the same; looking at the parentheses, is p = .05 the same as p < .05? That is what I am going off of…I ignored everything except the parentheses, you would need the underline under the left thing*
MI10T	correct		no	*I don't know what error rates are….so I will agree*
MI14T	correct	practical interest		
MI1F	correct		understands that this is wrong, but doesn't actually understand why	*I am forgetting what the null hypothesis is, but I do understand the question because you are basically asking me what the definition of a p-value is*
MI17F	incorrect	confirmatory		*I feel like the word 'confirmatory' is a trick word…maybe this is a play on words…you don't technically 'confirm'*
MI2F	incorrect			*observed association confusing; what is frustrating me is that I am forgetting what the null hypothesis is and I feel like if I knew that…so you're not asking me for the definition of a p-value, you're asking me how the p-value relates to the null hypothesis?*
MI5F	incorrect			*I understand the question, not sure if right/wrong*
MI6F	correct			
MI5T	incorrect			
MI12T	incorrect	numerical precision of computation method	no	*not sure I understand this, I kind of get it, but I don't…No, they should not be reported as exact values because anytime I have read a research paper, they have been reported as p < .05, there are 3 ways to write it with stars…*

T/F	Selection	Problem	Understands Idea	Transcript Excerpts (Dana)
MI8T	correct			*When I read something like this, "large effects can…", I think 'it can' so I agree*
MI17T	correct			*so in the pizza example, you get 30 minutes in your sample and I get an hour?*
MI10F	correct	fluidity of sentence threw me off	no	*oh, so the parentheses is describing Type I error? Maybe a second sentence with the definition of Type I error..."I know what a p-value is and I know what a Type I error is…the p-value is the probability that you got your result due to random chance…I am disagreeing with that because a .05 p-value is not a reason to reject because of a Type I error"*
MI4F	correct			

P/CP	**Selection**	**Problem**	**Understands Idea**	**Transcript Excerpts (Dana)**
MI3cp	incorrect		no	*I understood the CP…so because I understood that one, I am going to choose P. Did you not understand P? Not as well, but not because it was not written well. I think I am getting tripped up between like the null hypothesis and my understanding a lack of rememberance.*
MI7p	incorrect			
MI8cp	correct			*"I understand both of them"*
MI9p	correct	combined with observations more extreme	no	*what does that mean?…"I think I understand"…is it that we are doing one study and we have multiple observations across that study? Had to explain single vs. cumulative probability*
MI15cp	correct		maybe?	*so in other words, you are testing the same thing…[after clarification]…oh all the evidence means the null is true…ok I understand both now that you clarified that*
MI16p	correct			*I understand both of these but I am not certain about it, I am just going to choose P*
MI18cp	incorrect		likely not, but no evidence	*asked for no explanation, indicated understanding*

RV3	Selection	Problem	Understands Idea	Transcript Excerpts (Dana)
V3A1cp	correct		unlikely, but had to explain that the T/F was context-free but the vignettes are not; in other words, instead of using a generic p-value, I was inserting that which was applicable to the current scenario	*so that? [didn't understand the 4.8% was the same as the p-value]…had to clarify that there was nothing special about 4.8 specifically other than that WAS the p-value in this instance (i.e. this is not what makes this true/false)*
V3A17p	incorrect		likely no	*Oh, I like that "being in agreement" (as opposed to 'confirmatory')*
V3A3cp	incorrect			*I understand both, going to go with one on the left*
V3A9p	incorrect		no, inconsistent with response to MI9p, guessing?	*needed to re-explain 'observations more extreme'*
V3B4cp	correct		maybe?	*Can you maybe elaborate? I am not sure…the truth of the null is not established? [I know p is not significant, you need to decide if the null is true or not] I am going to go with the one on the right*
V3B12p	incorrect		no	*I am limited in terms of knowing how to report an inequality, but I honestly do understand what you are asking…I'm just going to go with the one on the right*
V3B16cp	correct		maybe? States 'I understand'	*how are these different? [the last word] Maybe it would help if the word was underlined. You read something and you assume you know what it says, I read both words as the same word*
V3B15p	incorrect		no	*the one on the right was more clear to me, I am going to choose that one…I really don't understand either one*

RV5	Selection	Problem	Understands Idea	Transcript Excerpts (Dana)
V5A10cp	correct	long-run frequency	no	*I think this is a lack of understanding on my part, but I am not really sure what everything on the right is saying, but I do understand what the left is conveying, but I don't think the one on the left is correct based on my knowledge, so I am going to go with the one on the right. I think it is 'long-run frequency', I don't have knowledge of what that means.*
V5A7p	incorrect		no, because it is less than .05, I am going to go with the one on the right	*however, the way that the data are unusual…meaning? [just because I have a tiny p-value, does that mean that what I found is important? Are statistical significance/practical importance separate constructs that must be evaluated independently?]*
V5A11cp	correct		she is thinking of this purely mathematically: p = .00042 IS less than .05, but that is not what makes this correct/incorrect. She has not considered the statistical interpretation beyond satisfaction of the mathematical inequality (gets the right answer but for the wrong reason)	*It is interesting because I feel like both of these are right, but obviously they are not….I think I am understanding them correctly…oh, I didn't understand it…I thought I understood it but I didn't and now I do…I saw the word 'correct', I guess when I read it the second time...[certainly this number is < .05, but that is not what is questionable here; what is questionable is does reporting a single value or inequality convey the same meaning?]...actually, I did not understand it the way you explained it...because you said the word 'convey', now I understand it*
V5A18p	correct		maybe?	*Why are these tripping me up? Yeah, I would agree with that [needed clarification]*
V5A14cp	correct (guess)		no	*I understand both of them but I'm just honestly not sure about the answer*
V5B2cp	correct	24.36%	no, seems to be conflating what we are testing	*oh it is just written differently (p as percent); that p-value is too big so I am going to go with the answer on the right*
V5B5p	correct		maybe?	*[it is just this null hypothesis re-written for you] Ok, I understand, so it is like I understood the one on the right better and I don't think that is true, so I am going to go with the one on the left*
V5B6cp	incorrect			*I understand the one on the left and I am like 90% sure that it is true…I breezed through this one because I felt like it was right…so, I am honestly not sure*
V5B13cp	correct		yes	*So, I understand the one on the right…that makes sense to me…didn't understand the negation to be a negation; it is a property of the test, so I pick the one on the right*
V5B8p	correct		maybe?	*Ok, I understand, I have heard of effect size, so I am going to go with the one on the left*

T/F	Selection	Problem	Understands Idea	Transcript Excerpts (Mike)
MI16T	incorrect	"perfect agreement"?	maybe	*I am disagreeing because of the word "perfect"*
MI13T	correct			
MI11T	correct			
MI10T	correct			
MI14T	correct			
MI1F	incorrect		?	
MI17F	correct	"confirmatory"	maybe	*"I agree it is confirmatory but very rare to say yes for sure"; you did this twice, how sure are you that this 'thing' is actually true?; I am reading confirmatory as, yes this confirms things, it moves me further towards believing it*
MI2F	incorrect	"observed association"	maybe	*what does this phrase mean?; the phrase 'test statistic' would have helped him, but might not clarify it for all participants*
MI5F	incorrect		?	
MI6F	correct	no problem		
MI5T	incorrect (maybe misread canNOT)	"definitive conclusion"	yes	*If the study had a large sample, I still wouldn't feel great about the 'definitive conclusion'; this is more of a "study thing" not necessarily a p-value thing.*
MI12T	correct		yes	*"this is more of a philosophical thing", I agree…I like to have that information because a .045 and a .0045 are not the same and shouldn't be interpreted the same*
MI8T	correct			
MI17T	correct			
MI10F	correct	"IS 5%"	no	*To me, it would depend on what the p-value was; "like if the p-value WAS .05, then I would say the chance was 5%"; if the statement had said p = .05, then I would have said yes (thinks p-value = prob. of Type I error); I might have gotten it correct on a technicality*
MI4F	correct			

P/CP	Selection	Understands Idea	Problem	Transcript Excerpts (Mike)
MI3cp	correct			
MI7p	correct	yes		*"that is one of my favorite ones", that something can be statistically significant, but not practically important*
MI8cp	correct			
MI9p	incorrect			
MI15cp	correct			
MI16p	correct	yes	"opposite sides of .05"	*".055 and .045 are the same result basically"*
MI18cp				

RV3	Selection	Problem	Understands Idea	Transcript Excerpts (Mike)
V3A1cp	incorrect		matches misinterpretation shown in MI1F	
V3A17p	correct			
V3A3cp	correct			
V3A9p	correct			
V3B4cp	correct			
V3B12p	correct			
V3B16cp	correct			
V3B15p	correct		yes	*I think it is fine, but I had to read it 3 times to be sure what it was talking about. Totality of evidence means to put the 2 studies together and look at as one big study.*

RV5	Selection	Problem	Understands Idea	Transcript Excerpts (Mike)
V5A10cp	incorrect			
V5A7p	correct			
V5A11cp	correct			
V5A18p	correct	"at least as small"	likely	*I would expect it to be below .05, but not necessarily AS small; "kind of a 50-50 chance, really"*
V5A14cp	correct	"should"	yes	*I said yes because I would want the one-sided because of the hypothesis, not sure what the standard is*
V5B2cp	incorrect		matches misinterpretation shown in MI2F	
V5B5p	correct			
V5B6cp	correct			*wordy, but I think it is fine; took me a little to wrap my mind around it*
V5B13cp	correct			
V5B8p	correct			

T/F	Selection	Problem	Understands Idea	Transcript Excerpts (Shelley)
MI16T	correct			*so by opposite sides, you mean one < .05 and one > .05; I understand the question*
MI13T	correct			*so, it seems to me that I disagree, if I understand correctly because it is a result of the test but also the effect and the population, all those things work together, so I think I disagree [is the test showing an effect, or does the effect live in the world] No, no, no...it is a property of the test, I agree. I thought you meant effect size and the population, but you didn't say sample, you said population. I was thinking the magnitude of the test value if you will plays into it also and the sample size, but it does specifically say population. [more clarification] It is a property of the test.*
MI11T	correct			*not sure I understand…if you give a specific p (p = .03) is that different than saying p < .05? [do they convey the same information] Well, if your alpha value is .05, they mean the same thing. [they result in the same decision, but do they convey the same meaning?] Oh, I thought you were asking about reject/not reject...[rereads]...oh, I see....the negative wording of this question is tricky...I agree that it is incorrect*
MI10T	correct			*error rates (do you mean Type I and Type II?) Oh gosh, I don't know if I know the answer to that. I would assume…I am not sure I know the answer, but the question is very clear.*
MI14T	correct			*The word 'practical interest' helps because you decide on 1-sided vs. 2-sided based on what you are trying to show, so I think that is very clear.*
MI1F	correct			*Disagree, but statement is clear.*
MI17F	correct			*So, you are saying if you test the same hypothesis twice with different samples and you get p = .04, to interpret as confirmatory means….No, I don't think they are confirmatory…It might help to define in some way…there is a difference between the decision and the conclusion*
MI2F	incorrect			*question seems fine, do you mean like the alpha? Oh, the p that is calculated. What do you mean, for the null hypothesis? Yeah, that is fine. That's right, right? [draws a picture explaining the shaded region (alpha) and calls that p) I guess I don't understand what a p-value is then!*
MI5F	correct	in favor of		*If you have a large p-value, can you say that you accept the null or fail to reject? [those are different] Yes, if I just read this question, I would say agree. [do you think that you could accept the null?] No, definitely not. [mentions size of p-value as influential] No, that's not right. Maybe use those words; replace 'in favor' with 'accept the null'*
MI6F	correct			*I think this question is sort of asking the same thing as MI5F and I think this is more clear.*

T/F	Selection	Problem	Understands Idea	Transcript Excerpts (Shelley)
MI5T	correct			*are you talking about the null or the alternative? Again, I think the negative is confusing. I have to go back and read it again.*
MI12T	correct			*I agree, and I think that is what APA says to do, although I think journals do it differently.*
MI8T	correct			*By small study, do you mean small sample? I don't think I could tell you what it [i.e. noise in the data] means, but I think in context I understand what it means.*
MI17T	correct			*observed association, I think of it as a covariation of the things you are testing…no it seems clear to me because I think about the association between 2 variables in a chi-square is jumping out in my head because you use terms like 'observed' and 'expected' but I think the observed association is just is there a relationship between the...yeah someone could have a study that shows a positive correlation and someone else could have shown a negative correlation and the p-values could be the same [repeats her words]...that is perfectly clear to me.*
MI10F	correct			*No, you have to calculate that…I mean the question is clear. You have to calculate alpha and beta values and that is based on sample size and some other things I am forgetting*
MI4F	correct			*14 and 10 are more clearly worded than 9, but I think they are asking the same thing.*

P/CP	Selection	Problem	Understands Idea	Transcript Excerpts (Shelley)
MI3cp	correct			*[Reads P], I disagree so I am picking the other one. I don't like the word 'falsity'. I read that and my brain stopped working for a minute.*
MI7p	correct			*so, unusual means…what? Let me read the other one. Are we talking about the difference between practical significance and statistical significance. You can get a significant difference in pre/post scores but the student only improved by 2 points. The word 'unusual' is a bit off. In Point, you said what you meant. In CP, it took time to parse out the difference in the statements. Maybe use 'clinical interest'.*
MI8cp	correct			
MI9p	correct			*so, the probability of getting those same results if the null is true…I think it is this one because that is not right…more extreme meaning? [it is a tail probability] yeah, absolutely, that is correct. Now that I am reading both, this is the correct one. Making this a 2-part definition might make it easier.*
MI15cp	correct			*so if a bunch of studies test the same hypothesis and they all fail to reject the null, are you supposed to take that as evidence altogether that upholds the null…I would think upholds the null means that you accept the null….so I guess I take the CP because it depends on a lot of stuff.*
MI16p	correct			*I think the question is clear.*
MI18cp	correct		right answer for the wrong reason, but then gives example of disparate sample sizes	*at least as small…you mean, they will produce a p-value that is the same or smaller? I agree with the CP because of the 'at least' as small….you would hope to get similar results but I don't know if you can say 'at least' as small…[after discussion]…I think just because you have 2 studies testing the same hypothesis does not mean that you are going to get similar p-values or approximate or smaller than - particularly smaller than I find weird...so if you ask approximate, it might be a better way to say it*

RV2	Selection	Problem	Understands Idea	Transcript Excerpts (Shelley)
V2A1cp	correct			*No, that's not true. Yeah, it is that one (referring to CP)*
V2A4p	incorrect? I think she changed her answer so not sure if scored correctly.		I think she does understand this idea, in other words a non-significant p does not prove the null is true, but she is hung up on the wording?	*The truth is not established with a Non-significant results…the truth of it? I wouldn't say that was truth. Well, the average wasn't 15, I don't know, I will pick this one. [attempts to clarify] I understand the question. It doesn't prove anything. Look, I wouldn't say it would prove either. I don't know about the truth. You would conclude that the [the key word here is truth..did you prove it?] No, you didn't prove it.*
V2A5cp	correct			*See, now this I agree with (after reading 'point'). Oh no no no, except in relation to those hypotheses with smaller p-values….oh not in favor of, you are just finding evidence to refute the null so it is this one.*
V2A14cp	correct			*(reads 'point') Disagree.*
V2A9p	correct	typo - missing the 'is 28.8%' part of the 'point'		*I don't understand this one. What is it asking? [clarifies] Does the p-value tell you the probability…so not sure why these sentences don't make sense to me. [clarifies the distinction between the two statements - difference is the 'at least as extreme' bit] I don't know why when I reading these they don't seem like complete sentences to me [is missing the is 28% part] The 28% is telling us that we will get that 14.8 or more.*
V2B10cp	correct			*Type I is false positive…No. I guess I agree with that (counterpoint) since I don't agree with that (point). I thought you could calculate the probability of a Type I error.*
V2B11p	correct			*I mean, I think it is incorrect to interpret them as having the same meaning but I think that is how they are reported.*
V2B18cp	correct			
V2B16p	correct			*I think, I mean I don't know. It seems to me that they are close enough that they are corroborating…I don't think they are contradictory results. [discussion of sig vs. not, but similar difference in same direction]*
V2B17cp	correct			*To me it seems like you have more people now and your waittime got closer to 30 minutes, they are both about 1%, but I would say that they are not necessarily in agreement because you are trending the other way and you get more participants you might actually see that the waittime is a result of your sample since you only had so many customers. I would say you should NOT assume these results to be in agreement.*

RV8	Selection	Problem	Understands Idea	Transcript Excerpts (Shelley)
V8A6cp	correct	change wording to emphasize 'effect' in both P and CP statements		*No evidence of a table effect, I think I agree with that. Hold on, what is the difference…[have we shown the null to be false] oh yeah, no association…incompatible, yeah you can't say no effect [one of these says we saw no effect, the other says there is no effect] So, ok, yeah, I thought to myself that these are saying the same thing, it might have helped if the one on the left...could you say 'which implies no effect'?*
V8A8p	correct			*No, significance does not imply effect size.*
V8A12cp	correct			*Exact value is better.*
V8A15p	correct			*I think it is the first one, you should look at the totality of evidence. If you have two small samples and neither is significant, if you put them together, you might have a significant result.*
V8B2cp	correct			
V8B13p	correct			*It is this one, I like the way you said this here "not a property of the phenomenon". There was a question earlier, I think the word phenomenon makes it very clear.*
V8B3cp	correct			
V8B7p	correct			*See, again here you said practical importance. I like that. It made it very clear what you are talking about. When I look at this data, I am thinking, ok it is statistically significant, but it is the difference between 2.5 and 3.5 calls…that is not a lot. Doesn't seem that way to me.*

Abigail	Babs
Comments	**Comments**
In reference to PCP directions, "so, basically, which one is right?"	I find this extra bit after the directions (or an accurate interpretation) a little bit odd. I think I would prefer, "If you believe it is true, select 'I agree'; if not, select 'I disagree'; note that TRUE means it is an accurate…put it as a second sentence. A lot to take in at once.
in reference to 2A, she said 'obviously a contrived problem'. She thought the scenario seemed unrealistic - specifically that professors would claim exactly 15 times per semester.	I am clicking the 'COUNTERPOINT' because I think that one is true, but I needed to scroll back up. So if you weren't here, I would have scrolled back up…is that right? Am I picking the true one?
in two-study scenarios, doesn't like the prompt repeated.	p-value tells me how likely I will be wrong if I reject the null (this is alpha, not p)
In general, I don't care how many questions per page, but for that one where I had to refer back to the numbers, it was annoying to have to scroll. I wanted to be able to line up those values, maybe if there was a chart?	I liked them spaced out. The vignettes were not long, so it didn't take a lot of brain power to read through and consider them. You have to go back and keep re-reading and keep scrolling.
My answers to the decontextualized things informed how I responded to the vignettes. But I also think, based on our discussion, I saw definite parallels with some of the pieces. I think if I would have started with the vignettes, I would have pictured some of the things more naturally in terms of or I might have been able to understand ... As much as I could, I was trying to create actual contexts for myself in answering the T/F...I think in terms of what i have done in the course, it is usually contextualized. Like in the powerpoint, you will get the decontextualized stuff and then you will have examples and you yourself will do examples and so refreshing with the examples first would be a little bit more natural because like remember this concept you did stuff with and now here is another example to apply it to, now in general what does this mean...so I guess, vignettes first.	I kind of like them how it was. The T/F kind of warmed up my brain. I am not sure if it would have made much difference. I would probably be too lazy to go back if a back button option were available. You always start with T/F, MC, and then problems - that is the order in my head so it is pleasing to me.
I like T/F before PCP. To some extent, it was short to long in that way.	Grad students don't care about extra credit. Don't they just get A's all the time?
I hate raffles. I feel like there is a very high probability that I will receive nothing so why did you tantalize me with the possibility?	

Dana	Mike	Shelley
Comments	**Comments**	**Comments**
Progress bar unnoticed. It would be nice to know the progress, but I would never have known what that was, even if I saw it. Would it help if pages with % were inserted? Yes, it would help. I'm like, how much more energy do I have to put into this?	Progress bar unnoticed	This one is definitely false, so I am going with the other one. Not sure if that is a good way to do it.
On some questions, I have to think about the selection…if I think the statement is false, does that mean I agree with the 'not' or do I disagree? The "nots" are throwing me off	too many items per page, but it didn't kill me	Doesn't like repeated scenarios in two-study vignette prompts.
I have to think and make sure that I am answering how I want to, ok, if this is NOT true, so if I think this is NOT true..I have to go back and make sure; maybe easier to parse if the not was not there	if it took 20 minutes, I would do it for a raffle if the prize was decent; if the time was 30-40 minutes, that would be the cut-off I would need to be paid; for coursework, I would do it up to an hour; longer than that, I would need to be paid	Can you at least put the part that is different in bold?
You go in thinking you are going to agree…so you are looking for evidence to disagree and if you don't find it, you will just choose agree?	I am taking this test off the top of my head, if I was reviewing, I would look stuff up. If I doubt myself, I would look. Now I might still be wrong if I don't doubt myself…but if there is any doubt, I am going to look it up (speaking to generalizability of results)	I don't like when you read a vignette and you spend all that time reading and then you only get asked 1 question.
True/false would have been easier than agree/disagree; I think correct/incorrect would have been easier, especially with the ones that had the 'not' I had to think about what is the correct answer and then what is the correct selection	In RV section, repeat directions at the top of every page	
forgot the direction after reading the first scenario, because the layout was the same, I assumed it was the same as P/CP, but I wasn't sure	Doesn't like "statistical distortion", not saying it is wrong, but hard to swallow; statistical distortion means the numbers came out distorting, the researcher isn't doing something, but the numbers came out wrong; rewrite to make it clear that the researcher was involved in this error; "the researchers distorted the statistical results", the statistics themselves did not distort	
sometimes if I don't understand the one, it helps to have the other one		
typo on MI9p		
T/F is a 'good warmup', RV are dense - don't want the hardest thing at the end		

Appendix O – PV2 and Item Modification Log

MI#	False Statement	True Statement	Point	Comment	Original Status
1	The p-value is the probability that the null hypothesis is true.	The p-value is NOT a measure of the probability of the truth of the null hypothesis.	TRUE	No change	F
2	The p-value for the null hypothesis is the probability that chance produced the observed association.	The probability that chance produced the observed association canNOT be determined from the p-value.	FALSE	No change	F
3	A significant test result ($p < 0.05$) means that the null hypothesis is false.	The falsity of the null hypothesis is NOT established with a significant result ($p \leq 0.05$)	TRUE	should inequality be strict? Inconsistent from P to CP; otherwise no change	P/CP
4	A nonsignificant test result ($p > 0.05$) means that the null hypothesis is true.	The truth of the null hypothesis is NOT established with a nonsignificant result ($p > 0.05$)	FALSE	No change, do we want strict inequality here?	F
5	A large p-value is evidence in favor of the null hypothesis.	A large p-value favors the null hypothesis only in relation to those hypotheses with smaller p-values.	TRUE	reworded to remove 'NOT'	T & F
6	A p-value > 0.05 means that the null hypothesis is incompatible with the observed data and therefore means that the study has demonstrated there to be "no association" or "no evidence" of an effect.	A p-value > 0.05 implies that the null hypothesis is incompatible with the observed data; however, unless the point estimate (observed association) equals the null value exactly, it is incorrect to conclude there to be "no association" or "no evidence" of an effect.	TRUE	reworded for parallelism	F
7	Statistical significance indicates a scientifically or substantively important relation has been detected.	Statistical significance indicates that the data are unusual in relation to the null hypothesis; however, the way that the data are unusual might not be scientifically or substantively important.	FALSE	clinical interest' replaced with 'scientifically or substantively important'	P/CP
8	Lack of statistical significance indicates that the effect size is small.	Large effect sizes can exist in the presence of insignificant p-values.	TRUE	reworded to remove 'NOT'	P/CP & T
9	The p-value is the probability of observing the data we observed if the null hypothesis is true.	The p-value is the probability of observing the data we observed, combined with observations more extreme than that which was observed, if the null hypothesis is true.	TRUE	No change	P/CP

MI#	False Statement	True Statement	Point	Comment	Original Status
10	If you reject the null hypothesis because p ≤ 0.05, the chance you have committed a Type I error in this instance (i.e., your "significant finding" is a false positive) is equal to the significance level of the test (i.e. α = 0.05).	The probability of committing a Type I error (i.e. your "significant finding" is a false positive) is a long-run frequency equivalent to the significance level of the test (i.e. α = 0.05) and not to the chance of error for any particular instance.	TRUE	reworded	T & F
11	It is correct to interpret the results of studies with reported values of p= 0.05 and p ≤ 0.05 as having the same meaning.	It is incorrect to interpret the results of studies with reported values of p= 0.05 and p ≤ 0.05 as having the same meaning.	FALSE	No change	T
12	P-values are properly reported as inequalities with familiar thresholds (i.e. p < .10, p < .05, p < .01) as opposed to exact values (i.e. p = .023).	P-values should be reported as exact values unless they are so small as to fall below the numerical precision of the computation method.	FALSE	No change	T
13	Statistical significance is a property of the phenomenon being studied which is detected/revealed by the statistical test.	Statistical significance is a property attached to the result of a statistical test.	TRUE	effect or population replaced with 'phenomenon' in TRUE option, statement shortened	T
14	When reporting results, two-sided p-values should always be used regardless of whether the hypothesis under investigation is one-sided or not.	When reporting results, the nature of the p-value (i.e. one-sided vs. two-sided) must be compatible with the hypothesis under investigation.	TRUE	‘hypothesis of practical interest' replaced with 'hypothesis under investigation'	T
15	When the same null hypothesis is tested in different studies and none (or a minority) of the tests are statistically significant, the overall evidence could show a statistically significant association or persuasive evidence of an effect.	When the same null hypothesis is tested in different studies and none (or a minority) of the tests are statistically significant, the overall evidence upholds the null hypothesis.	TRUE	‘totality of evidence' replaced with 'overall evidence', both stems reworded	P/CP
16	When the same hypothesis is tested twice, and the resulting p-values are on opposite sides of 0.05, the researcher should interpret these results as conflicting.	Two tests of the same hypothesis may produce p-values on opposite sides of 0.05 and yet the results may be in agreement.	FALSE	‘conflicting' replaces 'contradictory'; TRUE reworded to remove 'NOT'	P/CP & T
17	When the same hypothesis is tested twice and the resulting p-values are identical (or nearly so), the researcher should interpret these results as being in agreement.	Two tests of the same hypothesis may produce identical p-values and yet the results exhibit clearly different observed associations (i.e. conflicting results).	FALSE	‘confirmatory' replaced with 'in agreement', NOT removed	T & F
18	If one observes a small p-value, there is a good chance that the next study will produce a p-value at least as small for the same hypothesis.	If one observes a small p-value, there is no reason to believe that the next study has a good chance to produce a p-value at least as small for the same hypothesis.	TRUE	wording changed to remove 'NOT'	P/CP

Scenario	MI#	Point	CounterPoint	Point	CounterPoint
Scenario 3A: A manufacturer must test that his bolts are 2.00 cm long when they come off the assembly line. Bolts that are too long or too short cannot be used and the machines must be recalibrated in the event that unsatisfactory bolts are being produced. A random sample of 100 bolts yields a sample mean of 1.90 cm with a standard deviation of 0.5cm. The p-value for the appropriate test procedure was p = .048. H0: the average bolt length is 2.0 cm Ha: the average bolt length is not equal to 2.0 cm	1	There is a 4.8% chance that the average bolt length is 2.0 cm (i.e. the null is true).	The probability that the average bolt length is 2.0 cm (i.e. the null is true) cannot be determined from the p-value.	The chance that the average bolt length is 2.0 cm (i.e. the null is true) = .048.	The probability that the average bolt length is 2.0 cm (i.e. the null is true) cannot be determined from the p-value.
	17	Suppose a second random sample of 100 bolts (mean = 2.1 cm, standard deviation = 0.5 cm) yielded a p-value for the appropriate test procedure of p = .048. Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as being in agreement (i.e. confirmatory).	Suppose a second random sample of 100 bolts (mean = 2.1 cm, standard deviation = 0.5 cm) yielded a p-value for the appropriate test procedure of p = .048. Despite the fact that the resulting p-values are identical, it would be incorrect for the researcher to assume these results are in agreement.	(Scenario given in stem in lieu of repeating it). Despite the fact that the resulting p-values are identical for the same hypothesis, the researcher should NOT assume these results are in agreement.	Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as being in agreement.
	3	A significant test result (p = .048 < .05) means that the average bolt length is not equal to 2.0 cm (i.e. that the null is false).	The falsity of the null hypothesis (i.e. the average bolt length is not equal to 2.0 cm) is not established with a significant result (i.e. p = .048 < 0.05).	A significant test result (p = .048 < .05) means that the average bolt length is NOT equal to 2.0 cm (i.e. that the null is false).	The falsity of the null hypothesis (i.e. the average bolt length is not equal to 2.0 cm) is NOT established with a significant result (i.e. p = .048 < 0.05).
	9	The probability of observing the data we did (i.e. an average bolt length of 1.9 cm), if the null were true (i.e. average bolt length is 2.0 cm), is 4.8%.	The probability of observing data at least as extreme (i.e. an average bolt length ≤ 1.9 cm) as that which was observed, if the null were true (i.e. average bolt length is actually 2.0, is 4.8%.	The probability of observing data at least as extreme (i.e. an average bolt length &le; 1.9 cm) as that which was observed, if the null were true (i.e. average bolt length is actually 2.0), is .024.	The probability of observing the data we did (i.e. an average bolt length of 1.9 cm), if the null were true (i.e. average bolt length is 2.0 cm), is .048.

Scenario	MI#	Point	CounterPoint	Point	CounterPoint
Scenario 3B: A manufacturer is responsible for making barrels used to store crude oil, which are designed to hold exactly 55 gallons of oil. As part of a routine quality check, a factory manager randomly tests 27 barrels off the assembly line and finds the mean capacity is 54.7 gallons with a standard deviation of 0.8 gallons. The p-value for the appropriate test procedure was .062. H0: the average barrel capacity is 55 gallons Ha: the average barrel capacity is not equal to 55 gallons	4	A nonsignificant test result (p = .062 > 0.05) means that the average barrel capacity is 55 gallons (i.e. the null is true).	The truth of the null hypothesis (i.e. whether the average barrel capacity is 55 gallons) is not established with a non-significant result.	none	The truth of the null hypothesis (i.e. whether the average barrel capacity is 55 gallons) is NOT established with a non-significant result.
	12	This p-value is properly reported as an inequality: .05 < p < .10.	This p-value is properly reported as an exact value: p = .062.	This p-value is properly reported as an inequality: p < .10.	This p-value is properly reported as an exact value: p = .062.
	16	Suppose a second random sample of 35 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .033. Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the factory manager should interpret these results as contradictory/conflicting.	Suppose a second random sample of 35 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .033. Despite the fact that the resulting p-values are on opposite sides of 0.05, the researcher could interpret these results as confirmatory/corroborating.	(Scenario given in stem in lieu of repeating it). Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the factory manager should interpret these results as contradictory/conflicting.	Despite the fact that the resulting p-values are on opposite sides of 0.05, the researcher could interpret these results as confirmatory/corroborating.
	15	Suppose a second random sample of 25 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .073. Since the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager should conclude that the totality of evidence upholds the null hypothesis (i.e. the claim that the average barrel capacity is 55 gallons).	Suppose a second random sample of 25 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) yielded a p-value for the appropriate test procedure of p = .073. Although the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager cannot assume that the totality of evidence upholds the null hypothesis (i.e. the claim that the average barrel capacity is 55 gallons) and must consider that the studies taken collectively might lead to an entirely different conclusion than that of any individual result.	(Scenario given in stem in lieu of repeating it). Since the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager should conclude that the overall evidence upholds the null hypothesis (i.e. the claim that the average barrel capacity is 55 gallons).	Although the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager cannot assume that the overall evidence upholds the null hypothesis and must consider that the studies taken collectively might lead to an entirely different conclusion than that of any individual result.

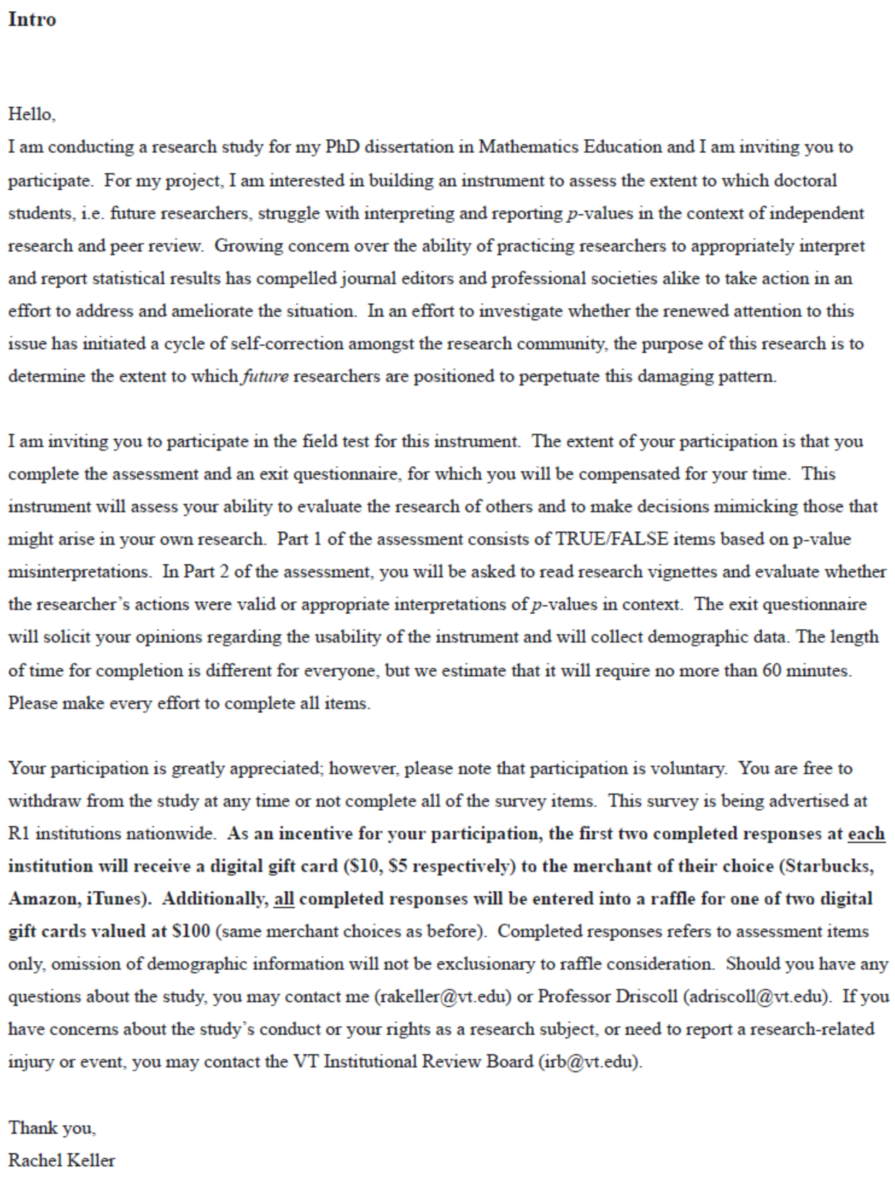
Scenario	MI#	Point	CounterPoint	Point	CounterPoint
Scenario 5A: A local school district is looking at adopting a new textbook that, according to the publishers, will increase standardized test scores of second graders by more than 10 points, on average. Never willing to believe a publisher's claim without evidence to support it, the school board decides to test the claim. Two classrooms are selected for the study, one of which is assigned the new textbook and the other used the traditional text. 8 children from each class were paired based on demographics and ability level resulting in a mean difference per pair of 11.5 points with standard deviation of .76 points. The (one-tailed) p-value for the appropriate test procedure was $p = .00042$. You may assume that all underlying assumptions have been satisfied and that there are no issues related to the experimental design (e.g. imprecise matching). H0: The difference between the groups will be ≤ 10 points. Ha: The difference between the groups will be > 10 points. .	10	The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) in this instance is 5%.	The probability of committing a Type I error (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) is a long-run frequency equivalent to the significance level of the test (i.e., $\alpha = 0.05$) and indeterminable for any particular hypothesis test.	The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error in this instance (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤ 10 points) is equal to the significance level of the test (i.e. $\alpha = 0.05$).	The probability of committing a Type I error (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) is a long-run frequency equivalent to the significance level of the test (i.e., $\alpha = 0.05$) and indeterminable for any particular hypothesis test.
	7	Statistical significance ($p = .00042 < .05$) indicates that a scientifically or substantively important relation between textbook and test scores has been detected.	Statistical significance ($p = .00042 < 0.05$) indicates that the observed data are unusual in relation to the null hypothesis (i.e. the new textbook yields score gains ≤ 10 points); however, the way that the data are unusual might be of no clinical interest or practical importance.	none	Statistical significance ($p = .00042 < 0.05$) indicates that the observed data (score difference = 11.5 points) are unusual in relation to the null hypothesis (i.e. the new textbook yields score gains ≤ 10 points); however, this might NOT be scientifically or substantively important.

Scenario	MI#	Point	CounterPoint	Point	CounterPoint
Scenario 5A H0: The difference between the groups will be ≤ 10 points. Ha: The difference between the groups will be > 10 points.	11	If one researcher had reported the results of this study as p < .05 and another reported p = .00042, it would be correct to interpret these representations as having the same meaning.	Reported p-values of p = .00042 and p < 0.05 do not convey the same meaning and should not be interpreted as equivalent.	If one researcher had reported the results of this study as p < .05 and another reported p = .00042, it would be correct to interpret these representations as having the same meaning.	Reported p-values of p = .00042 and p < 0.05 do NOT convey the same meaning and thus it would be incorrect to interpret them as equivalent.
	18	Since a small p-value was observed in this study (p = .00042), there is a good chance that the next study testing the same hypothesis (i.e. increase in test scores > 10 points) will produce a p-value at least as small.	Just because a small p-value was observed in this study (p = .00042), it cannot be assumed that the next study has a good chance to produce a p-value at least as small for the same hypothesis.	Since a small p-value was observed in this study (p = .00042), there is a good chance that the next study testing the same hypothesis (i.e. increase in test scores > 10 points) will produce a p-value at least as small.	Just because a small p-value was observed in this study (p = .00042), there is no reason to expect that the next study has a good chance to produce a p-value at least as small for the same hypothesis.
	14	This one-sided p-value (p = .00042) should be converted to a two-sided value when reporting the results of the study.	This one-sided p-value (p = .00042) should be reported as-is in order to maintain compatibility with the hypothesis of practical interest.	This one-sided p-value (p = .00042) should be converted to a two-sided value when reporting the results of the study.	This one-sided p-value (p = .00042) should be reported as-is in order to maintain compatibility with the hypothesis under investigation.

Scenario	MI#	Point	CounterPoint	Point	CounterPoint
Scenario 5B: An SAT prep course claims to increase scores by more than 60 points, on average. To test this claim, 9 students who have previously taken the SAT are randomly chosen to take the prep course. Their SAT scores before and after completing the prep course are compared and a mean difference of 81 points is found (standard deviation of 87 points). The (one-tailed) p-value for the appropriate test procedure was p = .2436. H0: Students who take the prep course improve their SAT scores by 60 points or less. Ha: Students who take the prep course improve their SAT scores by more than 60 points.	2	The probability that chance produced the observed association (i.e. improved scores by 81 points for those students in the prep course) is 24.36%.	The probability that chance produced the observed association between SAT scores and participation in the prep course) cannot be determined from the p-value.	The probability that chance produced the observed association (i.e. improved scores by 81 points for those students in the prep course) is equal to the p-value (p = .2436).	The probability that chance produced the observed association between SAT scores and participation in the prep course) cannot be determined from the p-value.
	5	Because this test produced a large p-value (p = .2436), this should be interpreted as evidence in favor of the claim that students who take the prep course improve their SAT score by 60 points or less	A large p-value cannot be said to favor the claim that the prep course leads to score improvements of ≤ 60 points (i.e. this specific null hypothesis) except in relation to those hypotheses with smaller p-values.	none	none
	6	Since our p-value was greater than 0.05, this means that the study has demonstrated there is no association between the prep course and score improvements greater than 60 points (in other words, the study has demonstrated, by virtue of the size of the p-value, that there is no evidence of a 60-point improvement effect attributable to the prep course).	A p-value > 0.05 (in this case, p = .2436) implies that the null hypothesis (i.e. score improvements of fewer than 60 points after prep course) is incompatible with the observed data (average improvement of 81 points); however, a conclusion of "no effect" (in other words, an effect size of zero) cannot be determined from this result.	Since our p-value was greater than 0.05, this means that the study has demonstrated there is no association between the prep course and score improvements greater than 60 points (in other words, the study has demonstrated, by virtue of the size of the p-value, that there is no evidence of a 60-point improvement effect attributable to the prep course).	A p-value > 0.05 (in this case, p = .2436) implies that the null hypothesis (i.e. score improvements of fewer than 60 points after prep course) is incompatible with the observed data (average improvement of 81 points); however, a conclusion of "no effect" (in other words, an effect size of zero) cannot be determined from this result.

Scenario	MI#	Point	CounterPoint	Point	CounterPoint
Scenario 5B H0: Students who take the prep course improve their SAT scores by 60 points or less. Ha: Students who take the prep course improve their SAT scores by more than 60 points.	13	Statistical significance is a property of the effect the prep course has on SAT scores and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.	Statistical significance is a property attached to the result of a statistical test and not a property of the phenomenon (i.e. the prep course's effect on SAT scores) under investigation.	Statistical significance is a property of the effect the prep course has on SAT scores and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.	Statistical significance is a property attached to the result of a statistical test and not a property of the phenomenon (i.e. the prep course's effect on SAT scores) under investigation.
	8	Lack of statistical significance (p = .2436 > .05) indicates that the effect size (i.e. the prep course's effect on SAT scores) is small.	The size of the effect is not determined by the significance/insignificance of the p-value; thus, a lack of statistical significance (p = .2436 > .05) does not necessarily imply that a small effect size has been detected.	none	It is possible that the prep course has a large effect on SAT scores despite the observed lack of statistical significance (p = .2436 > .05).

Appendix P – PV2 Qualtrics Version (Phase IIIA&B version)



Intro

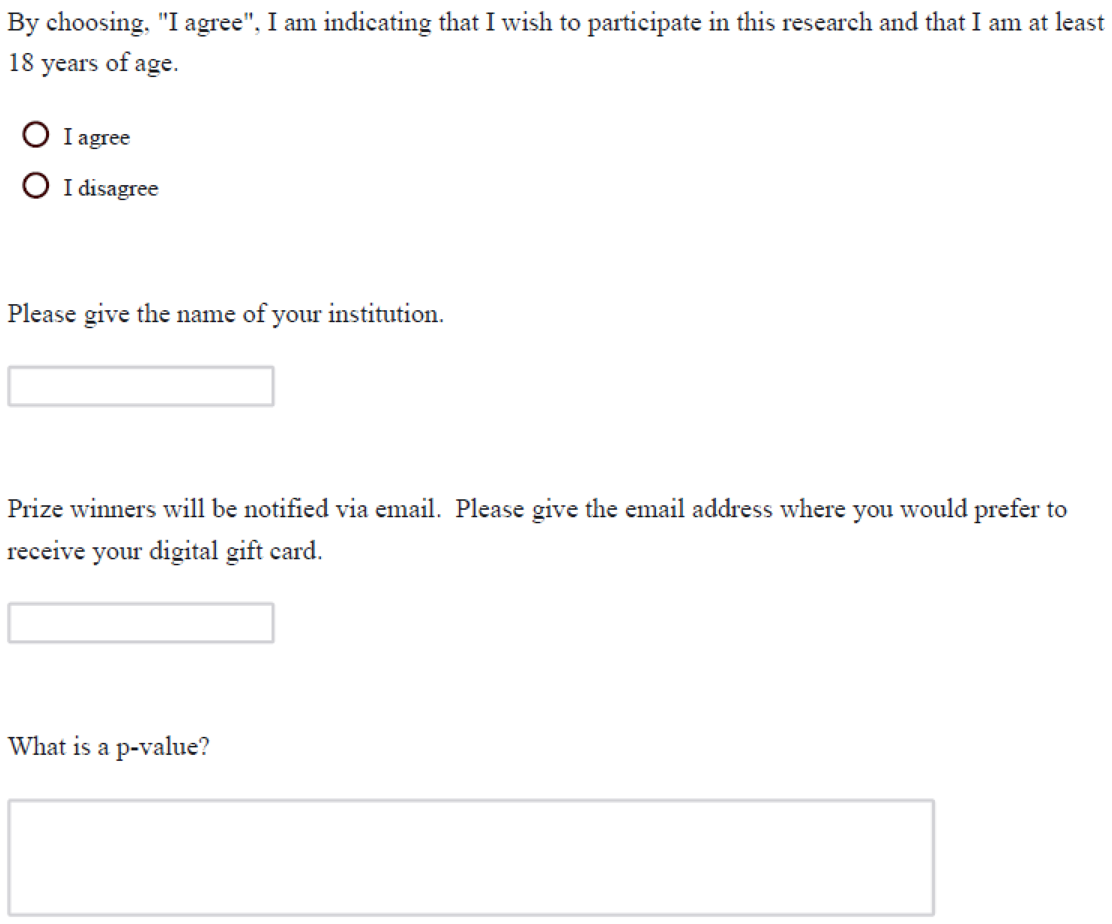
Hello,

I am conducting a research study for my PhD dissertation in Mathematics Education and I am inviting you to participate. For my project, I am interested in building an instrument to assess the extent to which doctoral students, i.e. future researchers, struggle with interpreting and reporting *p*-values in the context of independent research and peer review. Growing concern over the ability of practicing researchers to appropriately interpret and report statistical results has compelled journal editors and professional societies alike to take action in an effort to address and ameliorate the situation. In an effort to investigate whether the renewed attention to this issue has initiated a cycle of self-correction amongst the research community, the purpose of this research is to determine the extent to which *future* researchers are positioned to perpetuate this damaging pattern.

I am inviting you to participate in the field test for this instrument. The extent of your participation is that you complete the assessment and an exit questionnaire, for which you will be compensated for your time. This instrument will assess your ability to evaluate the research of others and to make decisions mimicking those that might arise in your own research. Part 1 of the assessment consists of TRUE/FALSE items based on p-value misinterpretations. In Part 2 of the assessment, you will be asked to read research vignettes and evaluate whether the researcher's actions were valid or appropriate interpretations of *p*-values in context. The exit questionnaire will solicit your opinions regarding the usability of the instrument and will collect demographic data. The length of time for completion is different for everyone, but we estimate that it will require no more than 60 minutes. Please make every effort to complete all items.

Your participation is greatly appreciated; however, please note that participation is voluntary. You are free to withdraw from the study at any time or not complete all of the survey items. This survey is being advertised at R1 institutions nationwide. **As an incentive for your participation, the first two completed responses at each institution will receive a digital gift card ($10, $5 respectively) to the merchant of their choice (Starbucks, Amazon, iTunes). Additionally, all completed responses will be entered into a raffle for one of two digital gift cards valued at $100** (same merchant choices as before). Completed responses refers to assessment items only, omission of demographic information will not be exclusionary to raffle consideration. Should you have any questions about the study, you may contact me (rakeller@vt.edu) or Professor Driscoll (adriscoll@vt.edu). If you have concerns about the study's conduct or your rights as a research subject, or need to report a research-related injury or event, you may contact the VT Institutional Review Board (irb@vt.edu).

Thank you,
Rachel Keller



By choosing, "I agree", I am indicating that I wish to participate in this research and that I am at least 18 years of age.

○ I agree

○ I disagree

Please give the name of your institution.

Prize winners will be notified via email. Please give the email address where you would prefer to receive your digital gift card.

What is a p-value?

PCP

PART I - POINT/COUNTERPOINT

For each of the following items, you will be presented with a pair of statements. Your task is to identify, within each pair, which one of the interpretations reflects an appropriate assertion or valid usage of p-values.

POINT

○ P-values are properly reported as inequalities with familiar thresholds (i.e. $p < .10$, $p < .05$, $p < .01$) as opposed to exact values (i.e. $p = .023$).

COUNTERPOINT

○ P-values should be reported as exact values unless they are so small as to fall below the numerical precision of the computation method.

POINT

○ The falsity of the null hypothesis is NOT established with a significant result ($p \leq 0.05$)

COUNTERPOINT

○ A significant test result ($p < 0.05$) means that the null hypothesis is false.

POINT

○ The p-value is NOT a measure of the probability of the truth of the null hypothesis.

COUNTERPOINT

○ The p-value is the probability that the null hypothesis is true.

POINT

○ Statistical significance is a property attached to the result of a statistical test.

COUNTERPOINT

○ Statistical significance is a property of the phenomenon being studied which is detected/revealed by the statistical test.

POINT

○ If one observes a small p-value, there is no reason to believe that the next study has a good chance to produce a p-value at least as small for the same hypothesis.

COUNTERPOINT

○ If one observes a small p-value, there is a good chance that the next study will produce a p-value at least as small for the same hypothesis.

POINT	COUNTERPOINT
○ The p-value is the probability of observing the data we observed, combined with observations more extreme than that which was observed, if the null hypothesis is true.	○ The p-value is the probability of observing the data we observed if the null hypothesis is true.

POINT	COUNTERPOINT
○ A nonsignificant test result ($p > 0.05$) means that the null hypothesis is true.	○ The truth of the null hypothesis is NOT established with a nonsignificant result ($p > 0.05$).

POINT	COUNTERPOINT
○ When the same hypothesis is tested twice, and the resulting p-values are on opposite sides of 0.05, the researcher should interpret these results as conflicting.	○ Two tests of the same hypothesis may produce p-values on opposite sides of 0.05 and yet the results may be in agreement.

POINT	COUNTERPOINT
○ When reporting results, the nature of the p-value (i.e. one-sided vs. two-sided) must be compatible with the hypothesis under investigation.	○ When reporting results, two-sided p-values should always be used regardless of whether the hypothesis under investigation is one-sided or not.

POINT	COUNTERPOINT
○ A p-value > 0.05 implies that the null hypothesis is incompatible with the observed data; however, unless the point estimate (observed association) equals the null value exactly, it is incorrect to conclude there to be "no association" or "no evidence" of an effect.	○ A p-value > 0.05 implies that the null hypothesis is incompatible with the observed data and therefore means that the study has demonstrated there to be "no association" or "no evidence" of an effect.

POINT	COUNTERPOINT
○ It is correct to interpret the results of studies with reported values of p= 0.05 and $p \leq 0.05$ as having the same meaning.	○ It is incorrect to interpret the results of studies with reported values of p= 0.05 and $p \leq 0.05$ as having the same meaning.

POINT	COUNTERPOINT
○ Statistical significance indicates a scientifically or substantively important relation has been detected.	○ Statistical significance indicates that the data are unusual in relation to the null hypothesis; however, the way that the data are unusual might NOT be scientifically or substantively important.

POINT

○ A large p-value favors the null hypothesis only in relation to those hypotheses with smaller p-values.

COUNTERPOINT

○ A large p-value is evidence in favor of the null hypothesis.

POINT

○ The p-value for the null hypothesis is the probability that chance produced the observed association.

COUNTERPOINT

○ The probability that chance produced the observed association cannot be determined from the p-value.

POINT

○ The probability of committing a Type I error (i.e.,your "significant finding" is a false positive) is a long-run frequency equivalent to the significance level of the test (i.e., $\alpha = 0.05$) and NOT to the chance of error for any particular instance.

COUNTERPOINT

○ If you reject the null hypothesis because $p \leq 0.05$, the chance you have committed a Type I error in this instance (i.e.,your "significant finding" is a false positive) is equal to the significance level of the test (i.e., $\alpha = 0.05$).

POINT

○ When the same null hypothesis is tested in different studies and none (or a minority) of the tests are statistically significant, the overall evidence could show a statistically significant association or persuasive evidence of an effect.

COUNTERPOINT

○ When the same null hypothesis is tested in different studies and none (or a minority) of the tests are statistically significant, the overall evidence upholds the null hypothesis.

POINT

○ Large effect sizes can exist in the presence of insignificant p-values.

COUNTERPOINT

○ Lack of statistical significance indicates that the effect size is small.

POINT

○ When the same hypothesis is tested twice and the resulting p-values are identical (or nearly so), the researcher should interpret these results as being in agreement.

COUNTERPOINT

○ Two tests of the same hypothesis may produce identical p-values and yet the results exhibit clearly different observed associations (i.e. conflicting results).

Part II - Research Vignettes

In the following section, you will be presented with a collection of research vignettes, each of which is accompanied by a series of paired statements meant to represent hypothetical interpretations of the research results. In the spirit of peer review, your task is to identify, within each pair which one of the interpretations reflects an appropriate assertion or valid usage of p-values.

Scenario 3A: A manufacturer must test that his bolts are 2.00 cm long when they come off the assembly line. Bolts that are too long or too short cannot be used and the machines must be recalibrated in the event that unsatisfactory bolts are being produced. A random sample of 100 bolts yields a sample mean of 1.90 cm with a standard deviation of 0.5 cm. The p-value for the appropriate test procedure was p = .048.

H0: the average bolt length is 2.0 cm
Ha: the average bolt length is not equal to 2.0 cm

POINT	COUNTERPOINT
○ The chance that the average bolt length is 2.0 cm (i.e. the null is true) = .048.	○ The probability that the average bolt length is 2.0 cm (i.e. the null is true) cannot be determined from the p-value.

POINT	COUNTERPOINT
○ A significant test result (p = .048 < .05) means that the average bolt length is NOT equal to 2.0 cm (i.e. that the null is false).	○ The falsity of the null hypothesis (i.e. the average bolt length is not equal to 2.0 cm) is NOT established with a significant result (i.e. p = .048 < 0.05).

POINT	COUNTERPOINT
○ The probability of observing data at least as extreme (i.e. an average bolt length ≤ 1.9 cm) as that which was observed, if the null were true (i.e. average bolt length is actually 2.0), is .024.	○ The probability of observing the data we did (i.e. an average bolt length of 1.9 cm), if the null were true (i.e. average bolt length is 2.0 cm), is .048.

Suppose a second random sample of 100 bolts (mean = 2.1 cm, standard deviation = 0.5 cm) yielded a p-value for the appropriate test procedure of p = .048.

POINT	COUNTERPOINT
○ Despite the fact that the resulting p-values are identical for the same hypothesis, the researcher should NOT assume these results are in agreement.	○ Since the same hypothesis was tested twice and the resulting p-values were identical, the manufacturer should interpret these results as being in agreement.

Scenario 3B: A manufacturer is responsible for making barrels used to store crude oil, which are designed to hold exactly 55 gallons of oil. As part of a routine quality check, a factory manager randomly tests 27 barrels off the assembly line and finds the mean capacity is 54.7 gallons with a standard deviation of 0.8 gallons. The p-value for the appropriate test procedure was .062.

H0: the average barrel capacity is 55 gallons
Ha: the average barrel capacity is not equal to 55 gallons

POINT

○ A nonsignificant test result ($p = .062 > 0.05$) means that the average barrel capacity is 55 gallons (i.e. the null is true).

COUNTERPOINT

○ The truth of the null hypothesis (i.e. whether the average barrel capacity is 55 gallons) is NOT established with a non-significant result.

POINT

○ This p-value is properly reported as an exact value: $p = .062$.

COUNTERPOINT

○ This p-value is properly reported as an inequality: $p < .10$.

Suppose instead that a second random sample of 25 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) had yielded a p-value for the appropriate test procedure of $p = .073$.

POINT

○ Since the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager should conclude that the overall evidence upholds the null hypothesis (i.e. the claim that the average barrel capacity is 55 gallons).

COUNTERPOINT

○ Although the same hypothesis was tested twice and neither of the tests was statistically significant, the factory manager cannot assume that the overall evidence upholds the null hypothesis and must consider that the studies taken collectively might lead to an entirely different conclusion than that of any individual result.

Suppose a second random sample of 35 barrels (mean = 54.7 gal, standard deviation = 0.8 gal) had yielded a p-value for the appropriate test procedure of $p = .033$.

POINT

○ Since the same hypothesis was tested twice and the resulting p-values are on opposite sides of 0.05, the factory manager should interpret these results as contradictory/conflicting.

COUNTERPOINT

○ Despite the fact that the resulting p-values are on opposite sides of 0.05, the researcher could interpret these results as confirmatory/corroborating.

Scenario 5A: A local school district is looking at adopting a new textbook that, according to the publishers, will increase standardized test scores of second graders by more than 10 points, on average. Never willing to believe a publisher's claim without evidence to support it, the school board decides to test the claim. Two classrooms are selected for the study, one of which is assigned the new textbook and the other used the traditional text. 8 children from each class were paired based on demographics and ability level resulting in a mean difference per pair of 11.5 points with standard deviation of .76 points. The (one-tailed) p-value for the appropriate test procedure was p = .00042. You may assume that all underlying assumptions have been satisfied and that there are no issues related to the experimental design (e.g. imprecise matching).

H0: The difference between the groups will be ≤ 10 points.
Ha: The difference between the groups will be > 10 points.

POINT

○ Statistical significance ($p = .00042 < 0.05$) indicates that the observed data (score difference = 11.5 points) are unusual in relation to the null hypothesis (i.e. the new textbook yields score gains ≤ 10 points); however, this might NOT be scientifically or substantively important.

COUNTERPOINT

○ Statistical significance ($p = .00042 < .05$) indicates that a scientifically or substantively important relation between textbook and test scores has been detected.

POINT

○ The null hypothesis was rejected since $p < 0.05$; therefore, the chance the researcher committed a Type I error in this instance (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) is equal to the significance level of the test (i.e., $\alpha = 0.05$).

COUNTERPOINT

○ The probability of committing a Type I error (i.e. rejecting the claim that the difference between the groups is 10 points or less when in fact the difference is ≤10 points) is a long-run frequency equivalent to the significance level of the test (i.e., $\alpha = 0.05$) and indeterminable for any particular hypothesis test.

POINT

○ If one researcher had reported the results of this study as $p < .05$ and another reported $p = .00042$, it would be correct to interpret these representations as having the same meaning.

COUNTERPOINT

○ Reported p-values of $p = .00042$ and $p < 0.05$ do NOT convey the same meaning and thus it would be incorrect to interpret them as equivalent.

POINT

○ This one-sided p-value ($p = .00042$) should be converted to a two-sided value when reporting the results of the study.

COUNTERPOINT

○ This one-sided p-value ($p = .00042$) should be reported as-is in order to maintain compatibility with the hypothesis under investigation.

POINT

○ Just because a small p-value was observed in this study (p = .00042), there is no reason to expect that the next study has a good chance to produce a p-value at least as small for the same hypothesis.

COUNTERPOINT

○ Since a small p-value was observed in this study (p = .00042), there is a good chance that the next study testing the same hypothesis (i.e. increase in test scores > 10 points) will produce a p-value at least as small.

Scenario 5B: An SAT prep course claims to increase scores by more than 60 points, on average. To test this claim, 9 students who have previously taken the SAT are randomly chosen to take the prep course. Their SAT scores before and after completing the prep course are compared and a mean difference of 81 points is found (standard deviation of 87 points). The (one-tailed) p-value for the appropriate test procedure was p = .2436.

H0: Students who take the prep course improve their SAT scores by 60 points or less.

Ha: Students who take the prep course improve their SAT scores by more than 60 points.

5B2

○ The probability that chance produced the observed association (i.e. improved scores by 81 points for those students in the prep course) is equal to the p-value (p = .2436).

○ The probability that chance produced the observed association between SAT scores and participation in the prep course) cannot be determined from the p-value.

5B5

○ A large p-value (p = .2436) only favors the claim that the prep course leads to score improvements of 60 points or less (i.e. this specific null hypothesis) in relation to those hypotheses with smaller p-values.

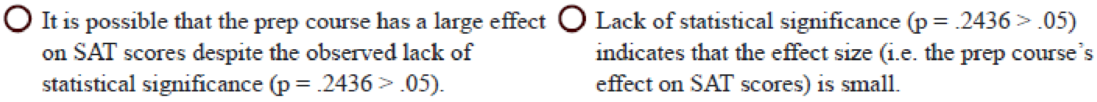
○ Because this test produced a large p-value (p = .2436), this should be interpreted as evidence in favor of the claim that students who take the prep course improve their SAT score by 60 points or less.

5B6

○ Since our p-value was greater than 0.05, this means that the study has demonstrated there is no association between the prep course and score improvements greater than 60 points (in other words, the study has demonstrated, by virtue of the size of the p-value, that there is no evidence of a 60-point improvement effect attributable to the prep course).

○ A p-value > 0.05 (in this case, p = .2436) implies that the null hypothesis (i.e. score improvements of fewer than 60 points after prep course) is incompatible with the observed data (average improvement of 81 points); however, a conclusion of "no effect" (in other words, an effect size of zero) cannot be determined from this result.

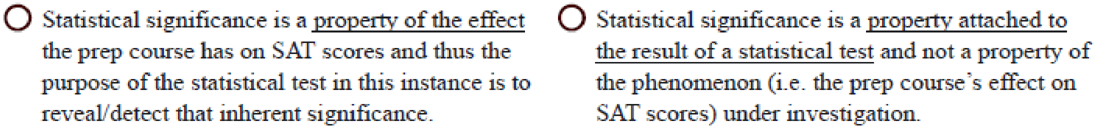
5B8



○ It is possible that the prep course has a large effect on SAT scores despite the observed lack of statistical significance (p = .2436 > .05).

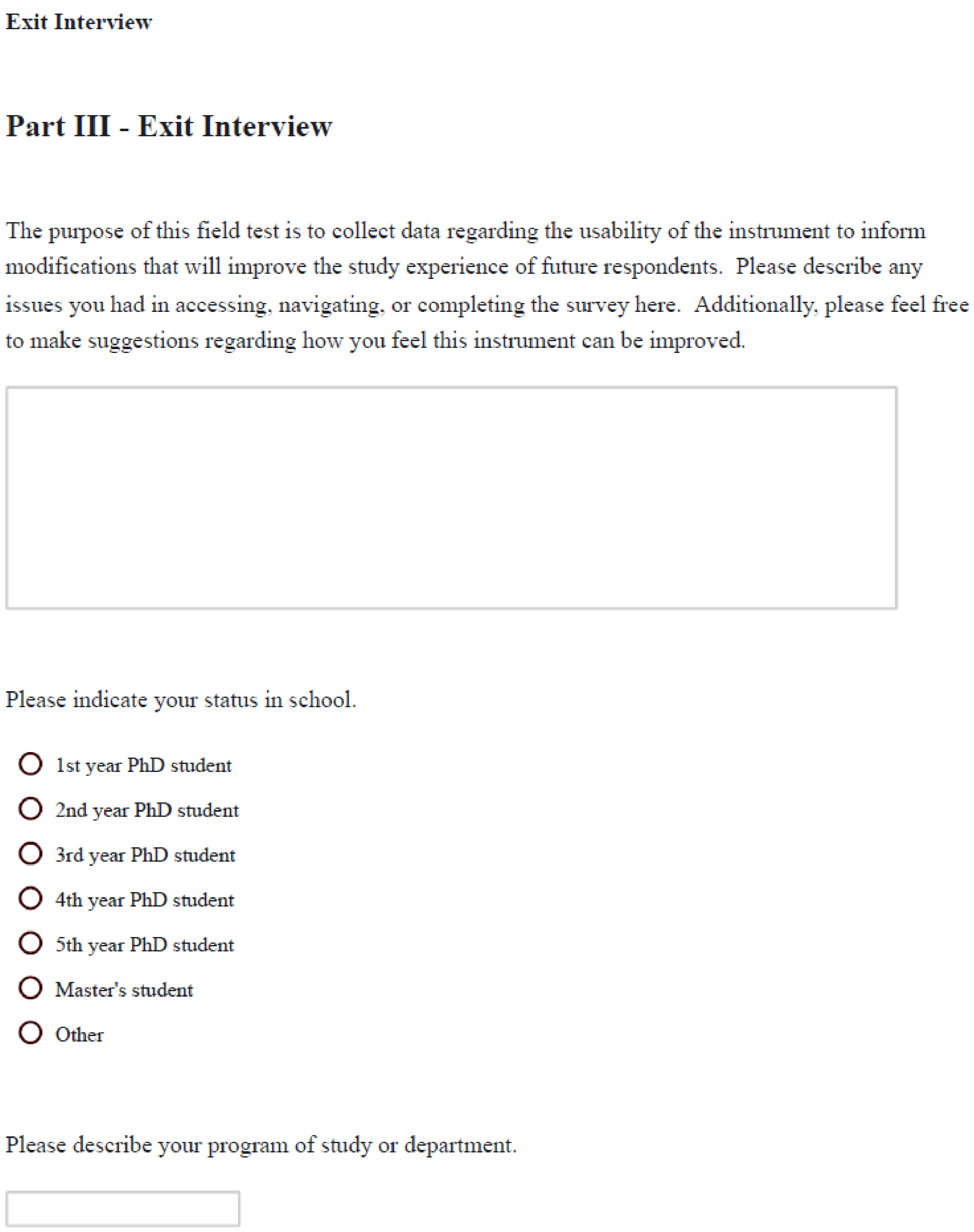
○ Lack of statistical significance (p = .2436 > .05) indicates that the effect size (i.e. the prep course's effect on SAT scores) is small.

5B13



○ Statistical significance is a property of the effect the prep course has on SAT scores and thus the purpose of the statistical test in this instance is to reveal/detect that inherent significance.

○ Statistical significance is a property attached to the result of a statistical test and not a property of the phenomenon (i.e. the prep course's effect on SAT scores) under investigation.



Exit Interview

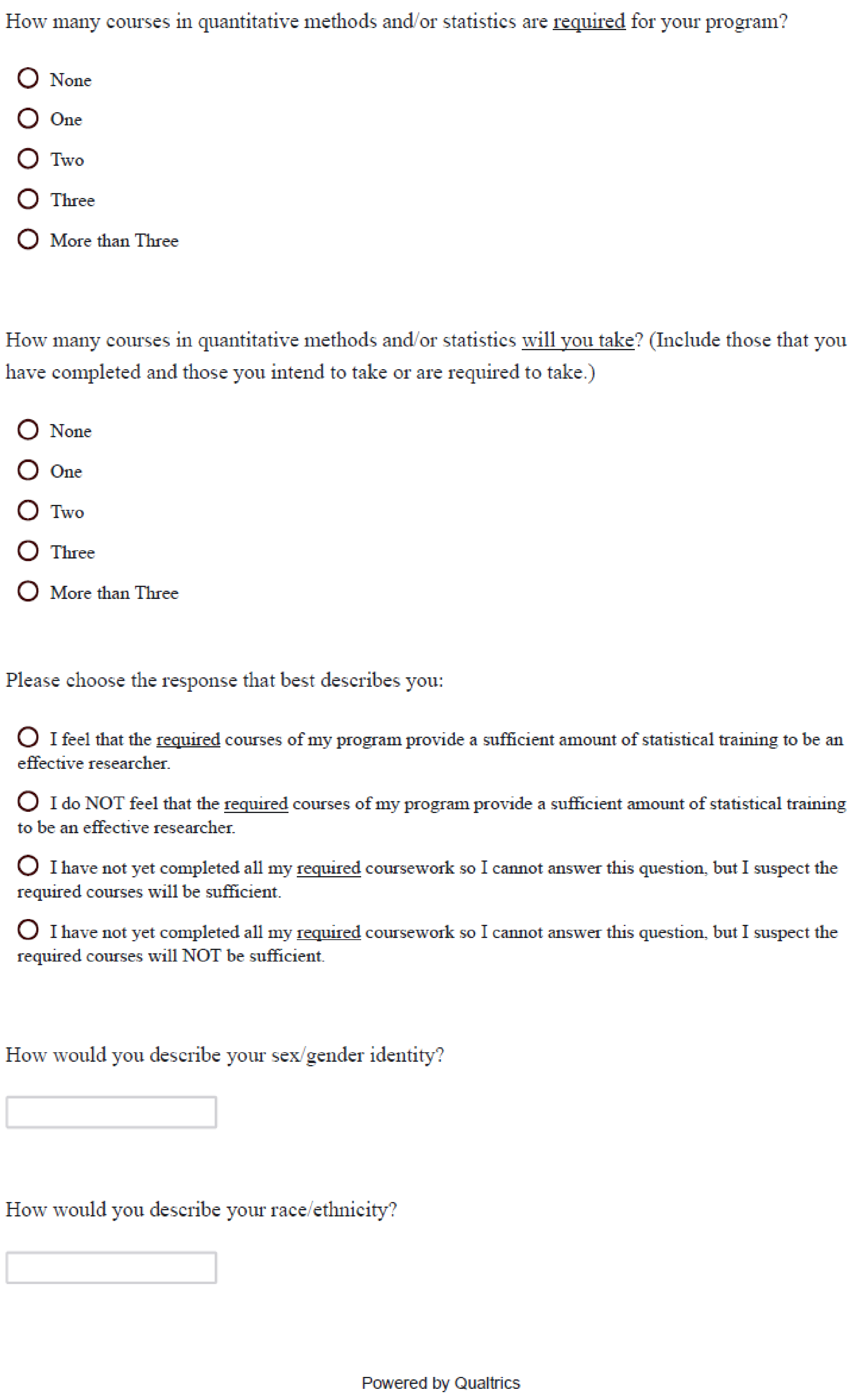
Part III - Exit Interview

The purpose of this field test is to collect data regarding the usability of the instrument to inform modifications that will improve the study experience of future respondents. Please describe any issues you had in accessing, navigating, or completing the survey here. Additionally, please feel free to make suggestions regarding how you feel this instrument can be improved.

Please indicate your status in school.

○ 1st year PhD student
○ 2nd year PhD student
○ 3rd year PhD student
○ 4th year PhD student
○ 5th year PhD student
○ Master's student
○ Other

Please describe your program of study or department.



How many courses in quantitative methods and/or statistics are required for your program?

○ None
○ One
○ Two
○ Three
○ More than Three

How many courses in quantitative methods and/or statistics will you take? (Include those that you have completed and those you intend to take or are required to take.)

○ None
○ One
○ Two
○ Three
○ More than Three

Please choose the response that best describes you:

○ I feel that the required courses of my program provide a sufficient amount of statistical training to be an effective researcher.

○ I do NOT feel that the required courses of my program provide a sufficient amount of statistical training to be an effective researcher.

○ I have not yet completed all my required coursework so I cannot answer this question, but I suspect the required courses will be sufficient.

○ I have not yet completed all my required coursework so I cannot answer this question, but I suspect the required courses will NOT be sufficient.

How would you describe your sex/gender identity?

□

How would you describe your race/ethnicity?

□

Powered by Qualtrics

Appendix Q – Phase IIIA Usability Comments

Some insight was gained regarding the reading and interpretation of p-values. However, I do not know if my responses were correct or incorrect. I'm having a hard time grasping how this is an instrument and how it will be utilized in the future.

A little bit more straightforward sentences in choices would be much appreciated.

Better to have more examples.

find a way how to make reading the text easier , maybe making bold some numbers. Mention the time this survey will take to complete,

From experience, sometimes it is better to approach a study like this from a simplistic perspective. The questions were needlessly long and verbose. I read everything, but in all honesty I wanted to tune out after a bit. This task is not mandatory nor graded so there is an inherent impulse not to put a large amount of effort into the answers. I would recommend one scenario that tests multiple facets of statistical learning.

I am not sure if this may serve the purpose of the instrument, but in the True/False of the first section, I felt like I wanted to answer Not sure/will review. A third Not sure answer may be needed. Some of the answers felt ambiguously similar such as in 5B5 and 5B6.

I believe the questions were repetitive in nature but that may have been on purpose to fully get a grasp on my understanding.

I completely didn't known the answer for some questions, and would liked if there was an "I don't know" option instead of me randomly checking a box.

I felt that this test was easy to access and navigate.

I had no issues navigating the survey. I would suggest adding something to show how far along you are in the survey (i.e. "Survey is 95% complete).

I left the survey for a couple hours and I was signed out of it when I wanted to finish it. As a result I had to redo it. It would be a good idea to have the responses saved as one goes through the survey.

I personally struggle with statistics, many of the concepts of the course are difficult to me because I do not see the math involved as tangible. Your quiz was good because it helped take away the math and got at the more important part- the interpretations.

I think it was very helpful for me. It improved me about how can I design the survey.

I think there were a few errors when writing statistically significant p-values (sometimes you wrote the p-value is significant but then wrote p-value&gt;0.05).

I think this is very important and should be used to help guide researchers in training on the best interpretation of "significance" in their future research.

I thought it was easy to navigate.

I thought the survey instructions were clearly defined and the questions were well-thought out.

I was slightly confused by the terminology used, but I do not think it was because it was poorly written. I think I was confused because I am not familiar with p-value questions. I am more so familiar with focusing on p-value being less than 0.05 and that representing significance.

I'm not the brightest with stats talk but I like to think I understand the underlying guidance of the test just taken. Some of the words tripped me up. My math speak isn't the best, nor is my english (and that's my first language). Also, I'm concerned with the lack of a third option on the two-choice questions. What if I had no clue what either of the statements meant and I had to guess (might have actually happened to me)? Seriously though, I just had a 30-minute conversation about that with someone. E-mail me your thoughts if you would to earleb@vt.edu, I was always taught in undergrad to leave an out for questions on knowledge-based surveys and am interested in your knowledge!

It didn't always feel like the points and counterpoints really related to one another.

It is hard to understand some of the question.. I guess can be improved with simple language.
Maybe in expressed in more simple way, a little bit difficult for me to understand the statement very clearly.
N/A
N/A
NA
No issues
No issues!
No issues, good luck with your study!
None
None
Not sure if its qualtrics or my computer browser (firefox) but each dot i clicked on took ~5 - 10 seconds to register the click. Even typeing in this box has taken a painfully long time for my words to show up after i type them
Put separate problems on a different page. It looks overwhelming.
Some of the questions are extremely wordy and I had a hard time following towards the end of the survey. Overall the survey was well put together and easy to understand.
The point a counter point sections wording sometimes gets a little confusing, also it has to be considered second language english speakers as part of the demographics
The survey instrument was well designed and the instructions were clear. I did not have trouble completing the survey.
The wording needs to be more clearer/basic
The words are too small to read.
Too many questions and very long. This experimental setting gets the individual bored after the first questions, and this might bias the results

Appendix R – Phase IIIA Reliability Data

P/CP Items

Case Processing Summary

	N	%
Valid	73	92.4
Excluded[a]	6	7.6
Total	79	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
0.328	18

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
9.4110	5.884	2.42576	18

Item Statistics

	Mean	Std. Deviation	N
M12	0.7534	0.43400	73
M3	0.2603	0.44182	73
M1	0.4110	0.49541	73
M13	0.4247	0.49771	73
M18	0.4932	0.50341	73
M9	0.5342	0.50228	73
M4	0.5890	0.49541	73
M16	0.5479	0.50114	73
M14	0.8630	0.34621	73
M6	0.5342	0.50228	73
M11	0.7534	0.43400	73
M7	0.5342	0.50228	73
M5	0.2466	0.43400	73
M2	0.5205	0.50303	73
M10	0.3151	0.46776	73
M15	0.5205	0.50303	73
M8	0.6986	0.46203	73
M17	0.4110	0.49541	73

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
M12	8.6575	5.756	-0.029	0.347
M3	9.1507	5.130	0.279	0.261
M1	9.0000	5.389	0.109	0.310
M13	8.9863	5.653	-0.007	0.345
M18	8.9178	5.882	-0.103	0.374
M9	8.8767	5.360	0.117	0.307
M4	8.8219	4.982	0.297	0.248
M16	8.8630	5.314	0.138	0.300
M14	8.5479	5.862	-0.058	0.348
M6	8.8767	5.443	0.081	0.318
M11	8.6575	5.617	0.038	0.329
M7	8.8767	5.443	0.081	0.318
M5	9.1644	5.056	0.328	0.248
M2	8.8904	5.849	-0.089	0.370
M10	9.0959	5.449	0.099	0.313
M15	8.8904	5.349	0.121	0.306
M8	8.7123	5.291	0.179	0.289
M17	9.0000	5.389	0.109	0.310

Odd Vignette Items

Case Processing Summary

		N	%
Cases	Valid	78	98.7
	Excluded[a]	1	1.3
	Total	79	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
0.511	18

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
9.5128	8.175	2.85922	18

Item Statistics

	Mean	Std. Deviation	N
V3A1	0.6282	0.48641	78
V3A3	0.3718	0.48641	78
V3A9	0.4615	0.50175	78
V3A17	0.5256	0.50257	78
V3B4	0.4615	0.50175	78
V3B12	0.8846	0.32155	78
V3B15	0.5769	0.49725	78
V3B16	0.3590	0.48280	78
V5A7	0.3974	0.49254	78
V5A10	0.4359	0.49908	78
V5A11	0.4615	0.50175	78
V5A14	0.7949	0.40641	78
V5A18	0.4872	0.50307	78
V5B2	0.5641	0.49908	78
V5B5	0.4872	0.50307	78
V5B6	0.5000	0.50324	78
V5B8	0.5641	0.49908	78
V5B13	0.5513	0.50058	78

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
V3A1	8.8846	7.376	0.213	0.487
V3A3	9.1410	7.577	0.135	0.503
V3A9	9.0513	8.257	-0.116	0.551
V3A17	8.9872	7.571	0.127	0.504
V3B4	9.0513	7.192	0.272	0.475
V3B12	8.6282	8.263	-0.103	0.532
V3B15	8.9359	6.944	0.375	0.453
V3B16	9.1538	7.223	0.277	0.475
V5A7	9.1154	7.584	0.129	0.504
V5A10	9.0769	7.397	0.195	0.491
V5A11	9.0513	7.062	0.323	0.464
V5A14	8.7179	8.413	-0.171	0.549
V5A18	9.0256	7.558	0.132	0.503
V5B2	8.9487	7.192	0.274	0.474
V5B5	9.0256	7.168	0.280	0.473
V5B6	9.0128	7.467	0.165	0.497
V5B8	8.9487	7.504	0.154	0.499
V5B13	8.9615	7.284	0.237	0.482

Full Instrument – All Items

Scale Statistics			
Mean	Variance	Std. Deviation	N of Items
19.0000	19.887	4.45952	36

Reliability Statistics	
Cronbach's Alpha	N of Items
0.599	36

Case Processing Summary

		N	%
Cases	Valid	72	91.1
	Excluded[a]	7	8.9
	Total	79	100.0

a. Listwise deletion based on all variables in the procedure.

Item Statistics			
	Mean	Std. Deviation	N
M12	0.7500	0.43605	72
M3	0.2639	0.44383	72
M1	0.4167	0.49647	72
M13	0.4167	0.49647	72
M18	0.5000	0.50351	72
M9	0.5278	0.50273	72
M4	0.5972	0.49390	72
M16	0.5556	0.50039	72
M14	0.8611	0.34826	72
M6	0.5417	0.50176	72
M11	0.7500	0.43605	72
M7	0.5417	0.50176	72
M5	0.2500	0.43605	72
M2	0.5139	0.50331	72
M10	0.3194	0.46953	72
M15	0.5278	0.50273	72
M8	0.7083	0.45772	72
M17	0.4167	0.49647	72
V3A1	0.6250	0.48752	72
V3A3	0.3750	0.48752	72
V3A9	0.4861	0.50331	72
V3A17	0.5139	0.50331	72
V3B4	0.4861	0.50331	72
V3B12	0.8750	0.33304	72
V3B15	0.5972	0.49390	72
V3B16	0.3333	0.47471	72
V5A7	0.4028	0.49390	72
V5A10	0.4583	0.50176	72
V5A11	0.4306	0.49863	72
V5A14	0.7778	0.41866	72
V5A18	0.4722	0.50273	72
V5B2	0.5556	0.50039	72
V5B5	0.5139	0.50331	72
V5B6	0.5139	0.50331	72
V5B8	0.5833	0.49647	72
V5B13	0.5417	0.50176	72

Item-Total Statistics				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
M12	18.2500	20.106	-0.104	0.614
M3	18.7361	18.591	0.287	0.580
M1	18.5833	19.486	0.035	0.603
M13	18.5833	18.951	0.160	0.591
M18	18.5000	19.352	0.064	0.601
M9	18.4722	19.408	0.051	0.602
M4	18.4028	18.497	0.270	0.580
M16	18.4444	18.476	0.270	0.580
M14	18.1389	19.980	-0.069	0.607
M6	18.4583	19.097	0.123	0.595
M11	18.2500	19.204	0.129	0.594
M7	18.4583	18.928	0.162	0.591
M5	18.7500	18.613	0.288	0.580
M2	18.4861	20.084	-0.100	0.616
M10	18.6806	19.263	0.098	0.597
M15	18.4722	18.450	0.274	0.580
M8	18.2917	19.111	0.142	0.593
M17	18.5833	18.725	0.213	0.586
V3A1	18.3750	18.576	0.256	0.582
V3A3	18.6250	18.914	0.174	0.590
V3A9	18.5139	19.972	-0.075	0.614
V3A17	18.4861	18.789	0.194	0.588
V3B4	18.5139	18.112	0.355	0.571
V3B12	18.1250	19.942	-0.056	0.606
V3B15	18.4028	18.300	0.318	0.576
V3B16	18.6667	18.197	0.362	0.572
V5A7	18.5972	18.864	0.182	0.589
V5A10	18.5417	19.322	0.071	0.600
V5A11	18.5694	18.418	0.285	0.579
V5A14	18.2222	20.260	-0.145	0.616
V5A18	18.5278	19.013	0.142	0.593
V5B2	18.4444	18.532	0.256	0.582
V5B5	18.4861	18.169	0.341	0.573
V5B6	18.4861	19.324	0.070	0.600
V5B8	18.4167	19.035	0.140	0.593
V5B13	18.4583	18.421	0.282	0.579

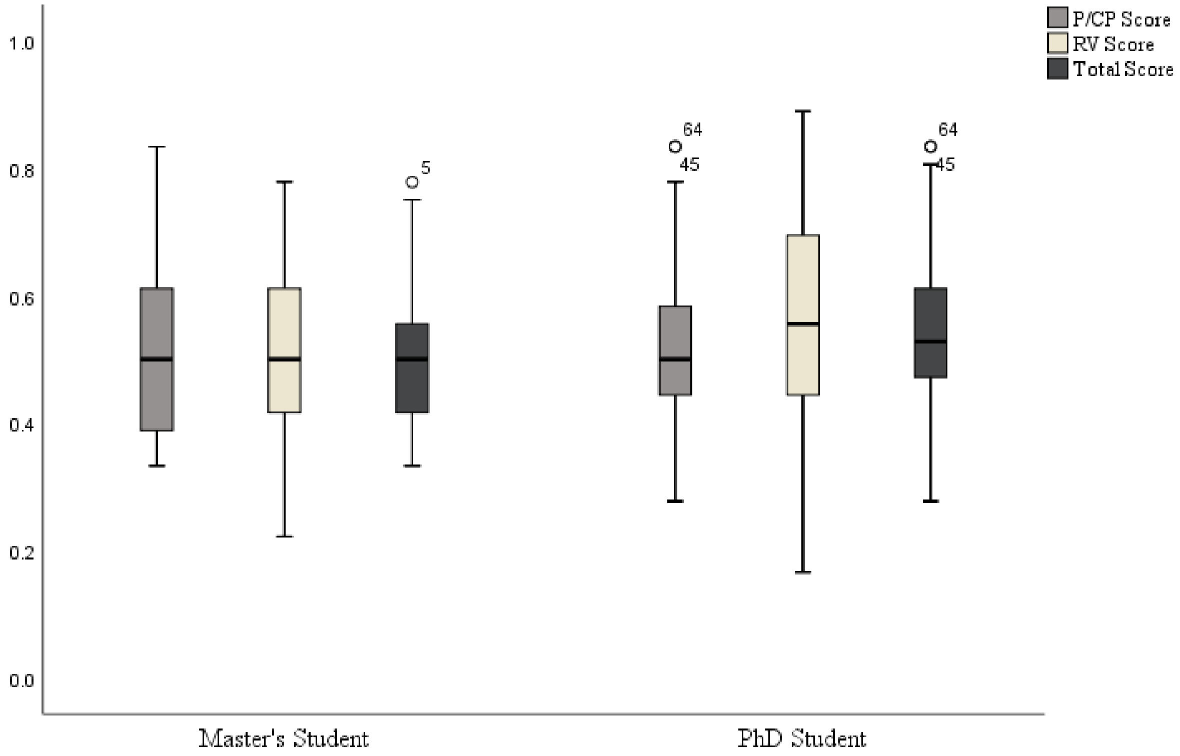
Appendix S – Phase IIIA Items Data by Population Subgroup

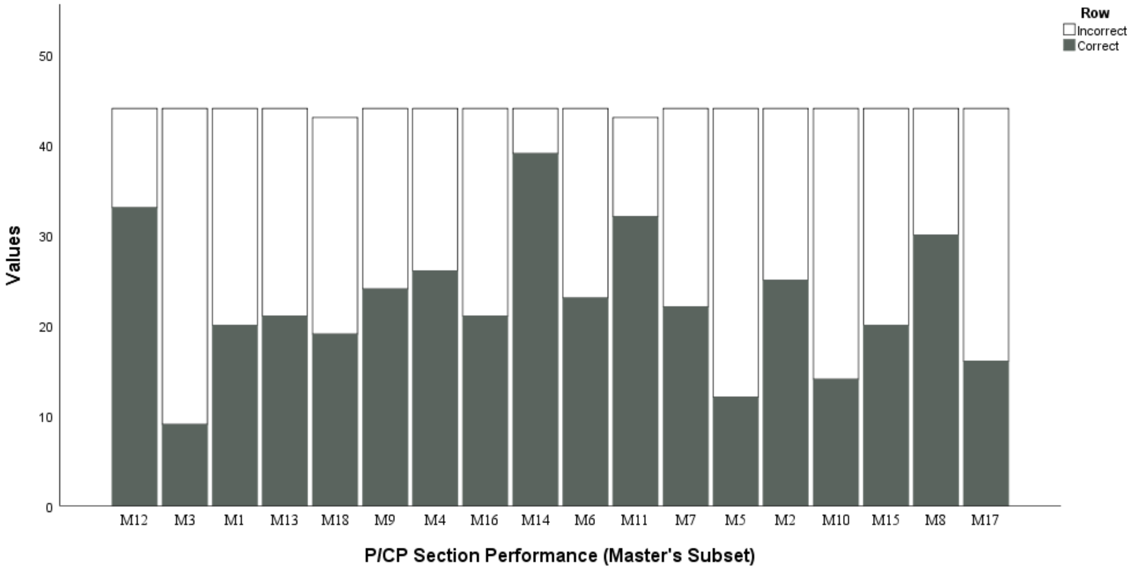
Descriptive Statistics - Master's Students

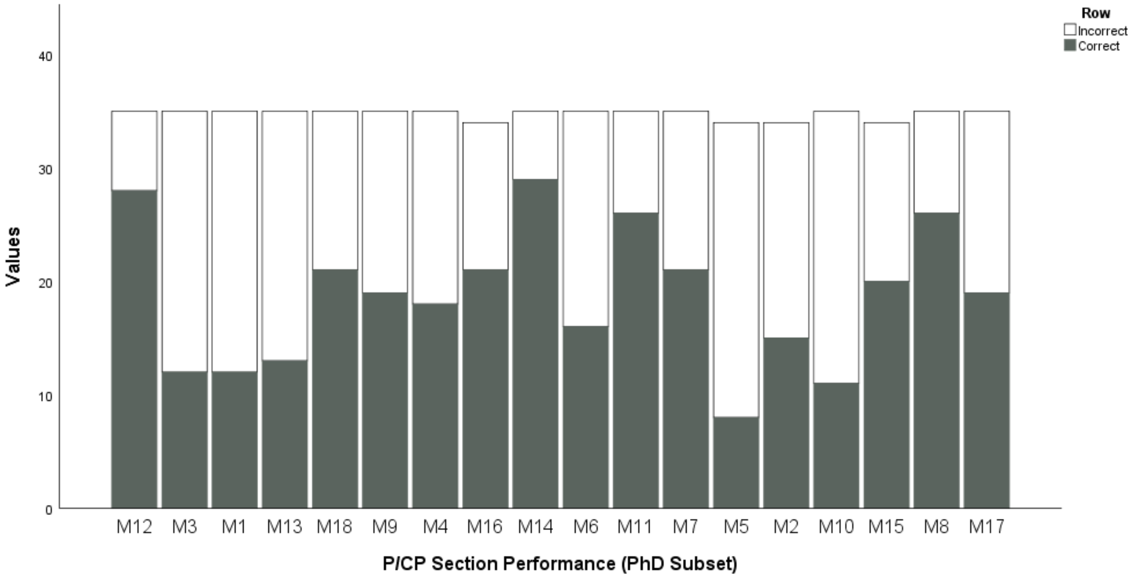
	N	Minimum	Maximum	Mean	Std. Deviation
P/CP Score	44	0.33	0.83	0.5126	0.12841
RV Score	44	0.22	0.78	0.5038	0.13092
Total Score	44	0.33	0.78	0.5082	0.10974
Valid N (listwise)	44				

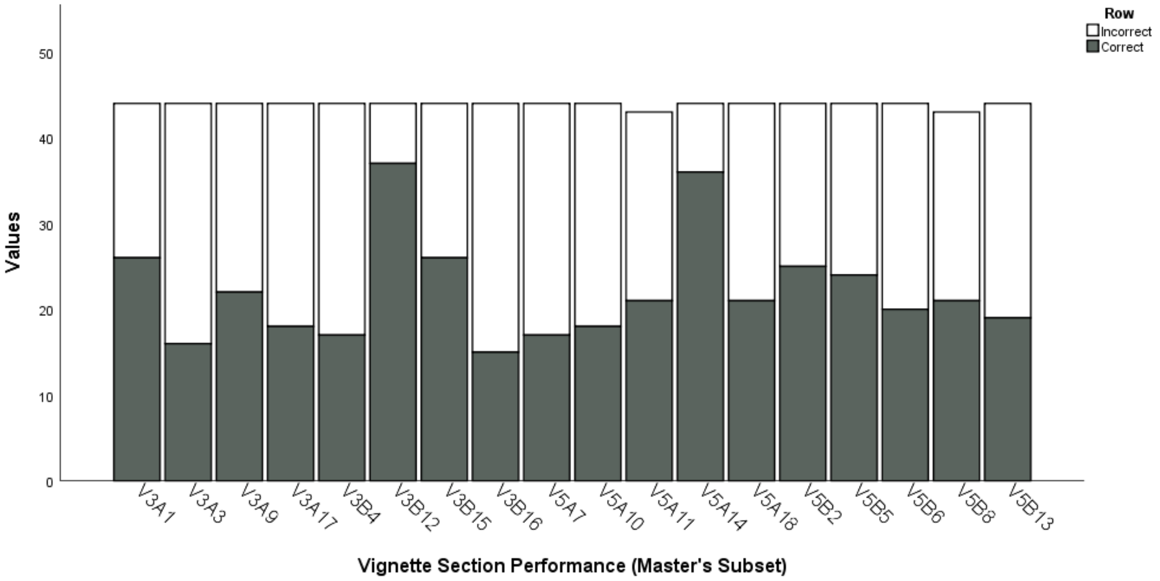
Descriptive Statistics - PhD Students

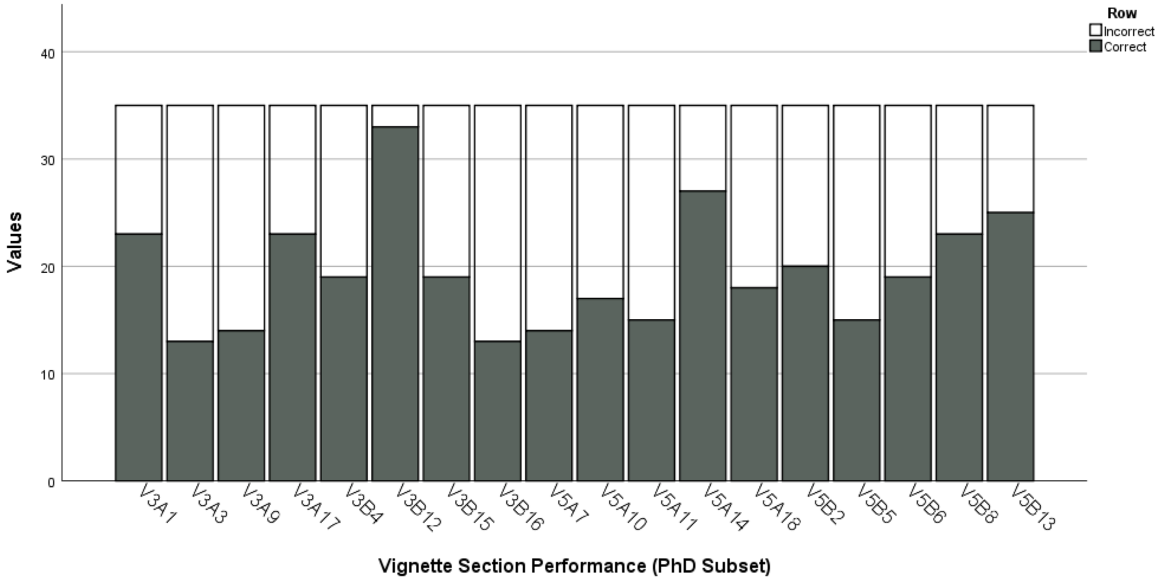
	N	Minimum	Maximum	Mean	Std. Deviation
P/CP Score	35	0.28	0.83	0.5317	0.14839
RV Score	35	0.17	0.89	0.5556	0.18573
Total Score	35	0.28	0.83	0.5437	0.14718
Valid N (listwise)	35				




Row
Incorrect
Correct
Values
0
10
20
30
40
50
M12
M3
M1
M13
M18
M9
M4
M16
M14
M6
M11
M7
M5
M2
M10
M15
M8
M17
P/CP Section Performance (Master's Subset)


Row
Incorrect
Correct
Values
0
10
20
30
40
M12
M3
M1
M13
M18
M9
M4
M16
M14
M6
M11
M7
M5
M2
M10
M15
M8
M17
P/CP Section Performance (PhD Subset)


Row
Incorrect
Correct
Values
0
10
20
30
40
50
V3A1
V3A3
V3A9
V3A17
V3B4
V3B12
V3B15
V3B16
V5A7
V5A10
V5A11
V5A14
V5A18
V5B2
V5B5
V5B6
V5B8
V5B13
Vignette Section Performance (Master's Subset)


Row
Incorrect
Correct
Values
0
10
20
30
40
V3A1
V3A3
V3A9
V3A17
V3B4
V3B12
V3B15
V3B16
V5A7
V5A10
V5A11
V5A14
V5A18
V5B2
V5B5
V5B6
V5B8
V5B13
Vignette Section Performance (PhD Subset)

Correlations - Master's Subset

		Score	P/CP Score	RV Score
Score	Pearson Correlation			
	Sig. (2-tailed)			
	N			
P/CP Score	Pearson Correlation	.843**		
	Sig. (2-tailed)	0.000		
	N	44		
RV Score	Pearson Correlation	.850**	.433**	
	Sig. (2-tailed)	0.000	0.003	
	N	44	44	
Taken Courses	Pearson Correlation	-0.235	-0.226	-0.172
	Sig. (2-tailed)	0.125	0.141	0.264
	N	44	44	44

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations - PhD Subset

		Score	P/CP Score	RV Score	Taken Courses
Score	Pearson Correlation				
	Sig. (2-tailed)				
	N				
P/CP Score	Pearson Correlation	.849**			
	Sig. (2-tailed)	0.000			
	N	35			
RV Score	Pearson Correlation	.907**	.547**		
	Sig. (2-tailed)	0.000	0.001		
	N	35	35		
Taken Courses	Pearson Correlation	.508**	.557**	.361*	
	Sig. (2-tailed)	0.002	0.001	0.033	
	N	35	35	35	
Status	Pearson Correlation	.354*	0.297	0.324	0.250
	Sig. (2-tailed)	0.037	0.083	0.058	0.147
	N	35	35	35	35

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Model Summary - Master's Subset

Model	R	R Square	Adj. R-Sq	Std. Error of the Estimate
1	.235[b]	0.055	0.033	0.10794

b. Predictors: (Constant), Taken Courses

ANOVA - Master's Subset

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	0.028	1	0.028	2.445	.125[c]
	Residual	0.489	42	0.012		
	Total	0.518	43			

b. Dependent Variable: Score2

c. Predictors: (Constant), Taken Courses

Coefficients - Master's Subset

Model		Unstd. Coefficients		Std. Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	0.566	0.040		14.029	0.000
	Taken Courses	-0.031	0.020	-0.235	-1.564	0.125

b. Dependent Variable: Score2

Model Summary - PhD Subset

Model	R	R Square	Adj. R-Sq	Std. Error of the Estimate
1	.508[b]	0.259	0.236	0.12864

b. Predictors: (Constant), Taken Courses

ANOVA - PhD Subset

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	0.190	1	0.190	11.507	.002[c]
	Residual	0.546	33	0.017		
	Total	0.737	34			

b. Dependent Variable: Score2

c. Predictors: (Constant), Taken Courses

Coefficients - PhD Subset

Model		Unstd. Coefficients		Std. Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	0.384	0.052		7.420	0.000
	Taken Courses	0.070	0.021	0.508	3.392	0.125

b. Dependent Variable: Score2

P/CP Reliability – Master’s Subset

Case Processing Summary - Master's Subset

		N	%
Cases	Valid	42	95.5
	Excluded[b]	2	4.5
	Total	44	100.0

b. Listwise deletion based on all variables in the procedure.

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
9.3333	5.350	2.31292	18

Reliability Statistics

Cronbach's Alpha	N of Items
0.253	18

Item Statistics - Master's Subset

	Mean	Std. Deviation	N
M12	0.7381	0.44500	42
M3	0.2143	0.41530	42
M1	0.4762	0.50549	42
M13	0.4762	0.50549	42
M18	0.4286	0.50087	42
M9	0.5476	0.50376	42
M4	0.6190	0.49151	42
M16	0.5000	0.50606	42
M14	0.8810	0.32777	42
M6	0.5476	0.50376	42
M11	0.7381	0.44500	42
M7	0.5000	0.50606	42
M5	0.2619	0.44500	42
M2	0.5952	0.49680	42
M10	0.3333	0.47712	42
M15	0.4524	0.50376	42
M8	0.6905	0.46790	42
M17	0.3333	0.47712	42

Item-Total Statistics - Master's Subset

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
M12	8.5952	5.027	0.062	0.244
M3	9.1190	5.083	0.050	0.247
M1	8.8571	5.052	0.018	0.260
M13	8.8571	5.101	-0.003	0.268
M18	8.9048	5.405	-0.132	0.311
M9	8.7857	5.099	-0.002	0.267
M4	8.7143	4.453	0.316	0.149
M16	8.8333	5.020	0.032	0.255
M14	8.4524	5.376	-0.088	0.279
M6	8.7857	4.953	0.064	0.243
M11	8.5952	5.222	-0.035	0.274
M7	8.8333	5.215	-0.053	0.285
M5	9.0714	4.263	0.484	0.097
M2	8.7381	4.881	0.101	0.230
M10	9.0000	4.927	0.092	0.233
M15	8.8810	4.742	0.161	0.207
M8	8.6429	4.772	0.176	0.205
M17	9.0000	5.073	0.023	0.257

P/CP Reliability – PhD Subset

Case Processing Summary - PhD Subset

		N	%
Cases	Valid	31	88.6
	Excluded[b]	4	11.4
	Total	35	100.0

b. Listwise deletion based on all variables in the procedure.

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
9.5161	6.791	2.60603	18

Reliability Statistics

Cronbach's Alpha	N of Items
0.424	18

Item Statistics - PhD Subset

	Mean	Std. Deviation	N
M12	0.7742	0.42502	31
M3	0.3226	0.47519	31
M1	0.3226	0.47519	31
M13	0.3548	0.48637	31
M18	0.5806	0.50161	31
M9	0.5161	0.50800	31
M4	0.5484	0.50588	31
M16	0.6129	0.49514	31
M14	0.8387	0.37388	31
M6	0.5161	0.50800	31
M11	0.7742	0.42502	31
M7	0.5806	0.50161	31
M5	0.2258	0.42502	31
M2	0.4194	0.50161	31
M10	0.2903	0.46141	31
M15	0.6129	0.49514	31
M8	0.7097	0.46141	31
M17	0.5161	0.50800	31

Item-Total Statistics - PhD Subset

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
M12	8.7419	6.931	-0.143	0.466
M3	9.1935	5.361	0.547	0.300
M1	9.1935	5.961	0.260	0.377
M13	9.1613	6.540	0.006	0.440
M18	8.9355	6.729	-0.073	0.459
M9	9.0000	5.867	0.271	0.372
M4	8.9677	5.832	0.288	0.367
M16	8.9032	5.890	0.273	0.372
M14	8.6774	6.692	-0.021	0.438
M6	9.0000	6.267	0.105	0.416
M11	8.7419	6.331	0.131	0.410
M7	8.9355	5.929	0.250	0.378
M5	9.2903	6.280	0.155	0.404
M2	9.0968	7.290	-0.277	0.506
M10	9.2258	6.314	0.114	0.413
M15	8.9032	6.357	0.076	0.423
M8	8.8065	6.161	0.182	0.397
M17	9.0000	6.000	0.214	0.387

RV Odd Reliability – Master’s Subset

Case Processing Summary - Master's Subset

		N	%
Cases	Valid	43	97.7
	Excluded[b]	1	2.3
	Total	44	100.0

b. Listwise deletion based on all variables in the procedure.

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
9.1163	5.581	2.36250	18

Reliability Statistics

Cronbach's Alpha	N of Items
0.248	18

Item Statistics - Master's Subset

	Mean	Std. Deviation	N
V3A1	0.6047	0.49471	43
V3A3	0.3721	0.48908	43
V3A9	0.5116	0.50578	43
V3A17	0.4186	0.49917	43
V3B4	0.3953	0.49471	43
V3B12	0.8372	0.37354	43
V3B15	0.6047	0.49471	43
V3B16	0.3488	0.48224	43
V5A7	0.3953	0.49471	43
V5A10	0.3953	0.49471	43
V5A11	0.4884	0.50578	43
V5A14	0.8140	0.39375	43
V5A18	0.4651	0.50468	43
V5B2	0.5581	0.50249	43
V5B5	0.5349	0.50468	43
V5B6	0.4651	0.50468	43
V5B8	0.4884	0.50578	43
V5B13	0.4186	0.49917	43

Item-Total Statistics - Master's Subset

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
V3A1	8.5116	5.018	0.144	0.210
V3A3	8.7442	5.195	0.066	0.238
V3A9	8.6047	5.388	-0.026	0.270
V3A17	8.6977	5.549	-0.092	0.292
V3B4	8.7209	4.539	0.378	0.120
V3B12	8.2791	5.682	-0.135	0.290
V3B15	8.5116	4.446	0.427	0.100
V3B16	8.7674	5.087	0.120	0.219
V5A7	8.7209	5.206	0.058	0.240
V5A10	8.7209	5.444	-0.047	0.276
V5A11	8.6279	4.811	0.232	0.175
V5A14	8.3023	5.835	-0.215	0.313
V5A18	8.6512	5.280	0.020	0.254
V5B2	8.5581	4.919	0.184	0.194
V5B5	8.5814	5.106	0.097	0.227
V5B6	8.6512	5.328	0.000	0.261
V5B8	8.6279	5.382	-0.024	0.270
V5B13	8.6977	5.359	-0.012	0.265

RV Odd Reliability – PhD Subset

Case Processing Summary - PhD Subset

		N	%
Cases	Valid	35	100.0
	Excluded[b]	0	0.0
	Total	35	100.0

b. Listwise deletion based on all variables in the procedure.

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
10.0000	11.176	3.34312	18

Reliability Statistics

Cronbach's Alpha	N of Items
0.665	18

Item Statistics - PhD Subset

	Mean	Std. Deviation	N
V3A1	0.6571	0.48159	35
V3A3	0.3714	0.49024	35
V3A9	0.4000	0.49705	35
V3A17	0.6571	0.48159	35
V3B4	0.5429	0.50543	35
V3B12	0.9429	0.23550	35
V3B15	0.5429	0.50543	35
V3B16	0.3714	0.49024	35
V5A7	0.4000	0.49705	35
V5A10	0.4857	0.50709	35
V5A11	0.4286	0.50210	35
V5A14	0.7714	0.42604	35
V5A18	0.5143	0.50709	35
V5B2	0.5714	0.50210	35
V5B5	0.4286	0.50210	35
V5B6	0.5429	0.50543	35
V5B8	0.6571	0.48159	35
V5B13	0.7143	0.45835	35

Item-Total Statistics - PhD Subset

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
V3A1	9.3429	10.114	0.271	0.650
V3A3	9.6286	10.299	0.202	0.659
V3A9	9.6000	11.482	-0.164	0.701
V3A17	9.3429	10.055	0.291	0.648
V3B4	9.4571	10.373	0.168	0.663
V3B12	9.0571	11.350	-0.144	0.679
V3B15	9.4571	9.726	0.379	0.637
V3B16	9.6286	9.652	0.422	0.632
V5A7	9.6000	10.306	0.195	0.660
V5A10	9.5143	9.669	0.397	0.634
V5A11	9.5714	9.546	0.444	0.628
V5A14	9.2286	11.358	-0.126	0.691
V5A18	9.4857	10.198	0.223	0.656
V5B2	9.4286	9.782	0.364	0.639
V5B5	9.5714	9.370	0.506	0.620
V5B6	9.4571	9.961	0.301	0.647
V5B8	9.3429	10.055	0.291	0.648
V5B13	9.2857	9.681	0.451	0.630

Full Instrument Reliability – Master’s Subset

Case Processing Summary - Master's Subset

		N	%
Cases	Valid	41	93.2
	Excluded[b]	3	6.8
	Total	44	100.0

b. Listwise deletion based on all variables in the procedure.

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
18.6098	14.944	3.86573	36

Reliability Statistics

Cronbach's Alpha	N of Items
0.454	36

Item Statistics - Master's Subset

	Mean	Std. Deviation	N		Mean	Std. Deviation	N
M12	0.7317	0.44857	41	V3A1	0.6098	0.49386	41
M3	0.2195	0.41906	41	V3A3	0.3659	0.48765	41
M1	0.4878	0.50606	41	V3A9	0.5366	0.50485	41
M13	0.4634	0.50485	41	V3A17	0.3902	0.49386	41
M18	0.4390	0.50243	41	V3B4	0.4146	0.49878	41
M9	0.5366	0.50485	41	V3B12	0.8293	0.38095	41
M4	0.6341	0.48765	41	V3B15	0.6341	0.48765	41
M16	0.5122	0.50606	41	V3B16	0.3171	0.47112	41
M14	0.8780	0.33129	41	V5A7	0.4146	0.49878	41
M6	0.5610	0.50243	41	V5A10	0.4146	0.49878	41
M11	0.7317	0.44857	41	V5A11	0.4634	0.50485	41
M7	0.5122	0.50606	41	V5A14	0.8049	0.40122	41
M5	0.2683	0.44857	41	V5A18	0.4634	0.50485	41
M2	0.5854	0.49878	41	V5B2	0.5610	0.50243	41
M10	0.3415	0.48009	41	V5B5	0.5610	0.50243	41
M15	0.4634	0.50485	41	V5B6	0.4878	0.50606	41
M8	0.7073	0.46065	41	V5B8	0.5122	0.50606	41
M17	0.3415	0.48009	41	V5B13	0.4146	0.49878	41

Item-Total Statistics - Master's Subset

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted		Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
M12	17.8780	14.910	-0.048	0.466	V3A1	18.0000	14.200	0.134	0.441
M3	18.3902	14.894	-0.039	0.464	V3A3	18.2439	14.539	0.045	0.455
M1	18.1220	15.010	-0.082	0.474	V3A9	18.0732	14.520	0.044	0.455
M13	18.1463	14.028	0.175	0.435	V3A17	18.2195	14.526	0.046	0.455
M18	18.1707	14.545	0.038	0.456	V3B4	18.1951	12.911	0.498	0.383
M9	18.0732	14.370	0.083	0.449	V3B12	17.7805	14.776	0.008	0.457
M4	17.9756	13.824	0.243	0.425	V3B15	17.9756	13.674	0.286	0.418
M16	18.0976	14.490	0.051	0.454	V3B16	18.2927	14.012	0.201	0.432
M14	17.7317	14.901	-0.026	0.460	V5A7	18.1951	14.261	0.115	0.444
M6	18.0488	13.948	0.198	0.431	V5A10	18.1951	15.061	-0.095	0.475
M11	17.8780	14.760	-0.005	0.461	V5A11	18.1463	13.878	0.216	0.429
M7	18.0976	14.640	0.012	0.460	V5A14	17.8049	15.211	-0.137	0.475
M5	18.3415	13.180	0.480	0.393	V5A18	18.1463	14.828	-0.036	0.467
M2	18.0244	14.224	0.125	0.443	V5B2	18.0488	14.248	0.117	0.444
M10	18.2683	13.951	0.213	0.430	V5B5	18.0488	13.948	0.198	0.431
M15	18.1463	13.728	0.257	0.422	V5B6	18.1220	14.760	-0.019	0.465
M8	17.9024	14.640	0.026	0.457	V5B8	18.0976	14.490	0.051	0.454
M17	18.2683	14.301	0.114	0.445	V5B13	18.1951	14.261	0.115	0.444

Full Instrument Reliability – PhD Subset

Case Processing Summary - PhD Subset

		N	%
Cases	Valid	31	88.6
	Excluded[b]	4	11.4
	Total	35	100.0

b. Listwise deletion based on all variables in the procedure.

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
19.5161	26.658	5.16314	36

Reliability Statistics

Cronbach's Alpha	N of Items
0.710	36

Item Statistics - PhD Subset

	Mean	Std. Deviation	N		Mean	Std. Deviation	N
M12	0.7742	0.42502	31	V3A1	0.6452	0.48637	31
M3	0.3226	0.47519	31	V3A3	0.3871	0.49514	31
M1	0.3226	0.47519	31	V3A9	0.4194	0.50161	31
M13	0.3548	0.48637	31	V3A17	0.6774	0.47519	31
M18	0.5806	0.50161	31	V3B4	0.5806	0.50161	31
M9	0.5161	0.50800	31	V3B12	0.9355	0.24973	31
M4	0.5484	0.50588	31	V3B15	0.5484	0.50588	31
M16	0.6129	0.49514	31	V3B16	0.3548	0.48637	31
M14	0.8387	0.37388	31	V5A7	0.3871	0.49514	31
M6	0.5161	0.50800	31	V5A10	0.5161	0.50800	31
M11	0.7742	0.42502	31	V5A11	0.3871	0.49514	31
M7	0.5806	0.50161	31	V5A14	0.7419	0.44480	31
M5	0.2258	0.42502	31	V5A18	0.4839	0.50800	31
M2	0.4194	0.50161	31	V5B2	0.5484	0.50588	31
M10	0.2903	0.46141	31	V5B5	0.4516	0.50588	31
M15	0.6129	0.49514	31	V5B6	0.5484	0.50588	31
M8	0.7097	0.46141	31	V5B8	0.6774	0.47519	31
M17	0.5161	0.50800	31	V5B13	0.7097	0.46141	31

Item-Total Statistics - PhD Subset

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted		Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
M12	18.7419	27.265	-0.177	0.725	V3A1	18.8710	24.583	0.381	0.694
M3	19.1935	23.761	0.577	0.682	V3A3	19.1290	24.916	0.303	0.699
M1	19.1935	25.428	0.210	0.705	V3A9	19.0968	27.290	-0.169	0.728
M13	19.1613	25.540	0.179	0.707	V3A17	18.8387	24.873	0.329	0.697
M18	18.9355	26.062	0.067	0.714	V3B4	18.9355	25.329	0.213	0.704
M9	19.0000	26.267	0.026	0.716	V3B12	18.5806	27.118	-0.201	0.719
M4	18.9677	24.766	0.325	0.697	V3B15	18.9677	24.499	0.380	0.694
M16	18.9032	24.024	0.492	0.687	V3B16	19.1613	23.940	0.521	0.685
M14	18.6774	26.892	-0.096	0.719	V5A7	19.1290	25.116	0.261	0.701
M6	19.0000	26.067	0.064	0.714	V5A10	19.0000	25.267	0.222	0.704
M11	18.7419	25.331	0.268	0.702	V5A11	19.1290	24.516	0.387	0.693
M7	18.9355	24.862	0.309	0.698	V5A14	18.7742	27.114	-0.141	0.724
M5	19.2903	25.946	0.123	0.709	V5A18	19.0323	24.766	0.323	0.697
M2	19.0968	27.890	-0.280	0.734	V5B2	18.9677	24.366	0.408	0.692
M10	19.2258	26.447	0.000	0.717	V5B5	19.0645	23.796	0.528	0.684
M15	18.9032	25.024	0.280	0.700	V5B6	18.9677	25.632	0.150	0.708
M8	18.8065	25.228	0.263	0.702	V5B8	18.8387	25.406	0.214	0.704
M17	19.0000	24.933	0.289	0.700	V5B13	18.8065	24.361	0.458	0.690

Appendix T – Phase IIIB Usability Comments

Each statement is too long.

First of all, you give away some answers with future questions. Secondly, I had a difficult determining when your questions referred to inferences I can make about the sample, versus inferences I can make about the population through the sample, which made some questions tricky. And yeah, I realized just after I clicked the "forward" button that I had chosen the wrong answer several times. Ah well. Good luck with your PhD!

Found myself over-analyzing the wording of the questions

I am a first-year student but the quarter hasn't started, so I've not yet had any graduate training in statistics--my responses reflect what I've learned in my undergraduate and independent studies.

I feel like the binary options generally implied a certain answer as being "correct" if it was more nuanced. In other words, I may have been less aware of misconceptions I held in a real peer-review situation than I was in this test. On the other hand, I would be interested to know if at any point I was accidentally mislead into the wrong answer by always assuming a more conservative answer was appropriate.

I found the wording in some of the questions to be a bit confusing.

I would have liked to see how well I did on the instrument so that I can correct any misunderstandings I may have.

I'm a 10th year PhD student: 2.5 yrs for a MSc in Physics, 2.5 yrs for another MSc in Planetary science, and 5 years with a new advisor; almost done! I've never needed to use p-values in my work. The closest I got was data fitting using SVD (and the stats therein). My wife has a PhD in Ed. Psych, and she is very statistically inclined. All I know about p-values comes from hearing her talk about her work. Bottom line is that I'm almost entirely clueless, and some of the terms (at least "effect size" and "long-run frequency") were essentially new to me. I took the survey, knowing I am incompetent in this field but not seeing any eligibility criteria I didn't fit. Hope this doesn't corrupt your sample! Feel free to fix my status as needed (e.g. 5th year PhD student). Statistical tests like these aren't really needed for my field, since the nature of the work we do is entirely different. Also, the vignette about the bolt factory has some issues with significant digits: the current value 2.00 cm necessarily implies a precision of 0.01 cm, yet elsewhere you report values with only one significant figure that imply a different precision. For example, you report a mean of 2.0 cm (rather than, e.g., 2.00 cm) with a standard deviation of 0.5 cm (rather than, e.g., 0.50 cm). This necessarily implies the precision of the observed measurements was 0.1 cm, which is less than the precision of the factory's requirement of 2.00 cm (within 0.01 cm). The nature of the study in the vignette wasn't clear to me as a result. In my field, how numbers are reported implies precision — in a similar way to how p-values imply meaning in stats. (The correct use of significant digits, rather than p-values, is more fundamental in my field of study.)

In general it felt like the questions had been written by a statistician who was interested in understanding how well I understood certain semantics of statistics. There were a few examples in which I wished I could have asked more questions about the situation to give a more nuanced response.   For example, I don't feel strongly about using $p < 0.01$ vs $p = 0.009$, I think either is appropriate and I am willing to use either equally. However, $p < 0.05$ makes me suspicious of why it is reported this way but $p = 0.0000000001$ isn't really necessary to convince me.

It is a great survey. Thank you!

It might be nice to cut the # of scenarios--the survey felt a bit long. It would also be nice to receive the answers at the end of the survey.

It might be nice to have a back button. After seeing some of the examples, I wanted to change some of my answers to the true/false questions.

It would be nice to know how many correct i got at the end of the study.

it's pretty systematic and good statistical practice.

Mechanically, this was a good survey. It was very frustrating, however, because it reminded me how poorly I understand p-values!

more pictures, it's hard to read all the text

N/A

N/A
n/a
N/A
n/a
No comments.
No Issues found
No issues.
No issues. Just noticed a few grammatical errors (e.g., "tests was" vs. "tests were").
No problems, but I did not know exactly what a "long-run frequency error" was and had to make an educated guess.
none
None
none. I thought it was easy to navigate and cleanly designed.
Nothing.
Once I read the vignettes, some of the questions in the first section became more clear and I would have changed my responses if I could go back (such as the conflicting findings with the same p-value with board length means that were greater than and subsequently less than the 2). I'm not sure who is in your target population, but if it's exclusively students in the social sciences I don't think it's as meaningful to write vignettes about studies that don't apply to the social sciences. For example, we would never read a study on board length; I think it would be much more relevant to give examples about latent variables (such as test scores, which was done later). Some of the True/False options were a little frustrating, because I wouldn't write about my findings in either way. For example, if the p-value is &gt; .05 we find no evidence of an effect, but that doesn't mean there is definitively no effect. But, I would never write my results section using the verbiage given; i.e. that the p-value is greater than .05 but that doesn't mean there is no effect. It would just confuse the reader. I would first conclude that this specific test found no effect, and then discuss other factors such as sample size, effect size, and the need to replicate. Regarding whether or not my program enables students to become an effective researcher... define effective researcher? Will students be able to be successful and do research? Yes, totally. Are they rigorously trained to understand the best methods and practices? No. Do I think that substantive researchers should be able to completely understand the best practices? Honestly, not really - I think that is why quant specialist counterparts exist, and I believe in collaboration. If the substantive researchers were quant experts, then we wouldn't have or need quant concentrations. I'm going to say no because I think you mean the latter, but I don't necessarily agree that they are therefore not effective researchers.
Some of the wordiness (particularly double-negatives) was/were a little difficult to wrap my brain around. The examples (part II) made the concepts seem more clear. I'm not sure if I did better on that part, but it certainly felt more intuitive.
survey access was fine; positive survey experience
The first page can be semantically confusing; it was easier to deal with the scenarios
The first section was lengthy. I think breaking it up into smaller sections may help keep participants focused on answering the questions
The format was nice, but I would have liked the ability to say "I don't know" at times. Selecting some answers was a coin toss to me, and I didn't like the idea of muddying your data that way.
The interface was great. The questions are difficult to parse at times.
The question about whether to report exact p-values or intervals is confusing and open to interpretation. A reformulation would be better. I believe that the question was meant to elicit the interpretation that the magnitude of the p-value matters or not, which I think it doesn't, only the arbitrary region the researcher chose. However, the interval region IS arbitrary, so I believe good practice is to report estimates and standard errors, so that the reader can easily calculate the p-value by hand, and indicatives of the region of rejection (ie, * p&lt;0.5) to help indicate the arbitrary region the researcher chose.

The questions were extremely redundant. If you are looking for congruency throughout the study then that seems fine, but all of the examples were pretty much the same asking the same true false questions. Good luck with your work.
The survey was easy to access, use and complete! Thank you!
The survey was easy to navigate. The questions were a little difficult, though. I took a stats class 3 years ago. I recognized a lot of the terminology from this survey but couldn't quite remember what it all meant. Statistics hasn't been directly relevant to my research thus far. (Aside from the required class I had to take.)
The wording on many of the questions was confusing. Additionally, there were some questions in which the null and research hypothesis were backwards making it difficult to answer. Null should be no finding.
There were no issues in accessing, navigating, or completing the survey that I could see. It worked smoothly, there was no trouble going to the next page or marking an answer to a question. There were no typos or grammatical errors. I often scrolled back to the top where the two hypotheses were written; if they could remain at the top of the screen while scrolling down the page it might help with efficiency.
Too hard, and I have not learned all of this yet. But it will be useful, I think.
Too long.
Using a Google pixel. It had opened in my email and I had accidently closed out of it when I had one come in. Would be better if could only open in browser so you couldn't accidently close out of it so easily.
Very interesting survey and study! I'm going to go look up some of these subtleties now. :)

Appendix U – Phase IIIB Reliability Data

P/CP Items

Case Processing Summary

	N	%
Valid	107	95.5
Excluded[a]	5	4.5
Total	112	100

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
0.509	18

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
10.99	7.613	2.759	18

Item Statistics

	Mean	Std. Deviation	N
M12	0.45	0.500	107
M3	0.59	0.494	107
M1	0.53	0.501	107
M13	0.64	0.481	107
M18	0.49	0.502	107
M9	0.42	0.496	107
M4	0.81	0.392	107
M16	0.64	0.484	107
M14	0.86	0.349	107
M6	0.53	0.501	107
M11	0.79	0.413	107
M7	0.85	0.358	107
M5	0.50	0.502	107
M2	0.48	0.502	107
M10	0.50	0.502	107
M15	0.51	0.502	107
M8	0.77	0.425	107
M17	0.64	0.481	107

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
M12	10.54	7.647	-0.103	0.548
M3	10.40	6.752	0.240	0.479
M1	10.46	6.534	0.323	0.460
M13	10.35	7.153	0.089	0.509
M18	10.50	7.290	0.026	0.523
M9	10.57	6.210	0.468	0.428
M4	10.18	6.827	0.309	0.471
M16	10.36	6.514	0.350	0.456
M14	10.13	7.473	0.010	0.518
M6	10.46	6.986	0.142	0.499
M11	10.21	7.335	0.048	0.514
M7	10.14	6.933	0.292	0.476
M5	10.50	6.800	0.214	0.484
M2	10.51	7.290	0.026	0.523
M10	10.50	7.290	0.026	0.523
M15	10.48	6.988	0.141	0.500
M8	10.22	6.987	0.198	0.489
M17	10.35	6.757	0.250	0.477

Odd Vignette Items

Case Processing Summary

	N	%
Valid	84	75.0
Excluded[a]	28	25.0
Total	112	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
0.608597882	18

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
11.06	9.237	3.039	18

Item Statistics

	Mean	Std. Deviation	N
V3A1cp	0.68	0.470	84
V3A3cp	0.44	0.499	84
V3A9p	0.40	0.494	84
V3A17p	0.79	0.413	84
V3B4cp	0.74	0.442	84
V3B12p	0.77	0.421	84
V3B15p	0.32	0.470	84
V3B16cp	0.55	0.501	84
V5A7p	0.57	0.498	84
V5A10cp	0.52	0.502	84
V5A11cp	0.50	0.503	84
V5A14cp	0.80	0.404	84
V5A18p	0.52	0.502	84
V5B2cp	0.56	0.499	84
V5B5p	0.71	0.454	84
V5B6cp	0.60	0.494	84
V5B8p	0.75	0.436	84
V5B13cp	0.83	0.375	84

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
V3A1cp	10.38	8.311	0.261	0.589
V3A3cp	10.62	8.287	0.244	0.591
V3A9p	10.65	7.988	0.360	0.572
V3A17p	10.27	8.225	0.355	0.577
V3B4cp	10.32	8.317	0.284	0.586
V3B12p	10.29	8.664	0.160	0.603
V3B15p	10.74	10.220	-0.401	0.677
V3B16cp	10.51	8.181	0.281	0.585
V5A7p	10.49	7.699	0.467	0.555
V5A10cp	10.54	8.372	0.211	0.596
V5A11cp	10.56	8.394	0.203	0.597
V5A14cp	10.26	8.629	0.187	0.599
V5A18p	10.54	8.541	0.151	0.605
V5B2cp	10.50	8.398	0.204	0.597
V5B5p	10.35	8.928	0.038	0.620
V5B6cp	10.46	8.035	0.342	0.575
V5B8p	10.31	8.240	0.323	0.581
V5B13cp	10.23	8.298	0.370	0.578

Full Instrument – All Items

Scale Statistics			
Mean	Variance	Std. Deviation	N of Items
22.01	26.240	5.123	36

Reliability Statistics	
Cronbach's Alpha	N of Items
0.718	36

Case Processing Summary

		N	%
Cases	Valid	80	71.4
	Excluded[a]	32	28.6
	Total	112	100.0

a. Listwise deletion based on all variables in the procedure.

Item Statistics

	Mean	Std. Deviation	N
M12	0.48	0.503	80
M3	0.58	0.497	80
M1	0.49	0.503	80
M13	0.61	0.490	80
M18	0.51	0.503	80
M9	0.36	0.484	80
M4	0.79	0.412	80
M16	0.61	0.490	80
M14	0.86	0.347	80
M6	0.53	0.503	80
M11	0.75	0.436	80
M7	0.85	0.359	80
M5	0.48	0.503	80
M2	0.50	0.503	80
M10	0.55	0.501	80
M15	0.56	0.499	80
M8	0.76	0.428	80
M17	0.65	0.480	80
V3A1	0.66	0.476	80
V3A3	0.45	0.501	80
V3A9	0.43	0.497	80
V3A17	0.79	0.412	80
V3B4	0.74	0.443	80
V3B12	0.78	0.420	80
V3B15	0.33	0.471	80
V3B16	0.54	0.502	80
V5A7	0.58	0.497	80
V5A10	0.53	0.503	80
V5A11	0.50	0.503	80
V5A14	0.80	0.403	80
V5A18	0.51	0.503	80
V5B2	0.58	0.497	80
V5B5	0.73	0.449	80
V5B6	0.59	0.495	80
V5B8	0.75	0.436	80
V5B13	0.85	0.359	80

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
M12	21.54	27.264	-0.243	0.740
M3	21.44	24.654	0.271	0.709
M1	21.53	24.202	0.361	0.703
M13	21.40	25.104	0.182	0.714
M18	21.50	25.848	0.027	0.724
M9	21.65	23.673	0.496	0.695
M4	21.23	24.328	0.429	0.701
M16	21.40	24.167	0.380	0.702
M14	21.15	25.673	0.127	0.717
M6	21.49	25.291	0.138	0.717
M11	21.26	25.994	0.013	0.723
M7	21.16	24.948	0.324	0.708
M5	21.54	24.961	0.205	0.713
M2	21.51	25.266	0.143	0.717
M10	21.46	25.568	0.083	0.720
M15	21.45	25.263	0.145	0.717
M8	21.25	24.671	0.326	0.706
M17	21.36	24.639	0.288	0.708
V3A1	21.35	24.509	0.319	0.706
V3A3	21.56	24.857	0.227	0.712
V3A9	21.59	24.347	0.335	0.705
V3A17	21.23	24.683	0.339	0.706
V3B4	21.28	24.354	0.387	0.703
V3B12	21.24	25.930	0.031	0.722
V3B15	21.69	27.762	-0.351	0.744
V3B16	21.48	24.961	0.205	0.713
V5A7	21.44	23.743	0.464	0.697
V5A10	21.49	24.506	0.298	0.707
V5A11	21.51	25.291	0.138	0.717
V5A14	21.21	25.334	0.184	0.714
V5A18	21.50	24.785	0.240	0.711
V5B2	21.44	24.502	0.303	0.707
V5B5	21.29	25.878	0.035	0.722
V5B6	21.43	24.197	0.369	0.703
V5B8	21.26	24.449	0.372	0.704
V5B13	21.16	24.492	0.455	0.702

Appendix V – jMetrik output for Classical Item Analysis by Sample

Item Analysis – VT Sample

Item	Option (Score)	Difficulty	Std. Dev.	Discrimin.
m12	Overall	0.8	0.4058	-0.2246
m3	Overall	0.3429	0.4816	0.7444
m1	Overall	0.3429	0.4816	0.3565
m13	Overall	0.3714	0.4902	0.2276
m18	Overall	0.6	0.4971	0.0801
m9	Overall	0.5429	0.5054	0.0341
m4	Overall	0.5143	0.5071	0.4713
m16	Overall	0.6	0.4971	0.6352
m14	Overall	0.8286	0.3824	-0.0975
m6	Overall	0.4571	0.5054	0.0441
m11	Overall	0.7429	0.4434	0.2149
m7	Overall	0.6	0.4971	0.3594
m5	Overall	0.2286	0.426	0.1143
m2	Overall	0.4286	0.5021	-0.1951
m10	Overall	0.3143	0.471	0.1276
m15	Overall	0.5714	0.5021	0.4191
m8	Overall	0.7429	0.4434	0.3205
m17	Overall	0.5429	0.5054	0.4067
v3a1	Overall	0.6571	0.4816	0.4469
v3a3	Overall	0.3714	0.4902	0.2879
v3a9	Overall	0.4	0.4971	-0.2293
v3a17	Overall	0.6571	0.4816	0.4788
v3b4	Overall	0.5429	0.5054	0.348
v3b12	Overall	0.9429	0.2355	-0.3669
v3b15	Overall	0.5429	0.5054	0.5258
v3b16	Overall	0.3714	0.4902	0.6779
v5a7	Overall	0.4	0.4971	0.3895
v5a10	Overall	0.4857	0.5071	0.362
v5a11	Overall	0.4286	0.5021	0.5071
v5a14	Overall	0.7714	0.426	-0.1724
v5a18	Overall	0.5143	0.5071	0.3683
v5b2	Overall	0.5714	0.5021	0.464
v5b5	Overall	0.4286	0.5021	0.5374
v5b6	Overall	0.5429	0.5054	0.2609
v5b8	Overall	0.6571	0.4816	0.3363
v5b13	Overall	0.7143	0.4583	0.528

```
=====================================================================
   TEST LEVEL STATISTICS – VT Sample
=====================================================================
Number of Items = 36
Number of Examinees = 35
Min = 10.0000
Max = 30.0000
Mean = 19.5714
Median = 19.0000
Standard Deviation = 5.2224
Interquartile Range = 5.0000
Skewness = 0.2045
Kurtosis = -0.1813
KR21 = 0.6917
======================================================================
```

```
                    RELIABILITY ANALYSIS – VT Sample
========================================================================
```

Method	Estimate	95% Conf. Int.		SEM
Guttman's L2	0.7651	(0.6380,	0.8639)	2.5681
Coefficient Alpha	0.7269	(0.5790,	0.8417)	2.7692
Feldt-Gilmer	0.7452	(0.6073,	0.8523)	2.6745
Feldt-Brennan	0.7363	(0.5936,	0.8472)	2.7209
Raju's Beta	0.7269	(0.5790,	0.8417)	2.7692

```
               RELIABILITY IF ITEM DELETED – VT Sample
=================================================================
```

Item	L2	Alpha	F-G	F-B	Raju
m12	0.7739	0.7389	0.7555	0.7476	0.7389
m3	0.7434	0.7007	0.7212	0.7103	0.7007
m1	0.7586	0.7182	0.7388	0.7285	0.7182
m13	0.763	0.7238	0.7435	0.7338	0.7238
m18	0.7678	0.7304	0.7481	0.7397	0.7304
m9	0.77	0.7326	0.7508	0.742	0.7326
m4	0.7528	0.7121	0.7321	0.7219	0.7121
m16	0.7463	0.7047	0.7242	0.7141	0.7047
m14	0.7704	0.7341	0.7512	0.7429	0.7341
m6	0.7696	0.7321	0.7496	0.7413	0.7321
m11	0.7629	0.7246	0.7429	0.7339	0.7246
m7	0.7575	0.7177	0.7378	0.7278	0.7177
m5	0.766	0.7282	0.7466	0.7377	0.7282
m2	0.7775	0.7426	0.7584	0.7512	0.7426
m10	0.7663	0.7281	0.7466	0.7377	0.7281
m15	0.7545	0.7148	0.7346	0.7246	0.7148
m8	0.7599	0.7206	0.7405	0.7307	0.7206
m17	0.7552	0.7152	0.736	0.7255	0.7152
v3a1	0.7552	0.7142	0.7347	0.7244	0.7142
v3a3	0.7602	0.7211	0.7403	0.7308	0.7211
v3a9	0.7782	0.7438	0.7589	0.7521	0.7438
v3a17	0.7534	0.7128	0.7327	0.7227	0.7128
v3b4	0.7575	0.718	0.7377	0.7277	0.718
v3b12	0.77	0.7337	0.7503	0.7422	0.7337
v3b15	0.7504	0.7095	0.7293	0.7192	0.7095
v3b16	0.7454	0.7031	0.7233	0.7128	0.7031
v5a7	0.7563	0.7163	0.7364	0.7263	0.7163
v5a10	0.7569	0.7173	0.7364	0.7269	0.7173
v5a11	0.7519	0.7106	0.7309	0.7206	0.7106
v5a14	0.7736	0.7381	0.7545	0.7467	0.7381
v5a18	0.7565	0.717	0.7369	0.727	0.717
v5b2	0.7538	0.7126	0.7334	0.7229	0.7126
v5b5	0.75	0.7091	0.7281	0.7184	0.7091
v5b6	0.7613	0.7221	0.7414	0.7319	0.7221
v5b8	0.7584	0.7191	0.7383	0.7288	0.7191
v5b13	0.7525	0.7118	0.7312	0.7214	0.7118

```
==============================================================
L2: Guttman's lambda-2
Alpha: Coefficient alpha
F-G: Feldt-Gilmer coeff.
F-B: Feldt-Brennan coef.
Raju: Raju's beta coeff.
```

Item Analysis – National Sample

Item	Option	Difficulty	Std. Dev.	Discrimin.
m12	Overall	0.4375	0.4983	-0.1878
m3	Overall	0.5982	0.4925	0.2646
m1	Overall	0.5268	0.5015	0.2184
m13	Overall	0.6339	0.4839	0.0721
m18	Overall	0.4821	0.5019	0.0814
m9	Overall	0.4196	0.4957	0.2719
m4	Overall	0.8125	0.3921	0.3764
m16	Overall	0.6518	0.4785	0.2541
m14	Overall	0.8661	0.3421	0.1487
m6	Overall	0.5268	0.5015	0.16
m11	Overall	0.7857	0.4122	-0.1513
m7	Overall	0.8482	0.3604	0.4314
m5	Overall	0.4911	0.5022	0.0692
m2	Overall	0.4732	0.5015	0.0976
m10	Overall	0.4821	0.5019	0.1391
m15	Overall	0.5089	0.5022	0.337
m8	Overall	0.7768	0.4183	0.3144
m17	Overall	0.6518	0.4785	0.2795
v3a1	Overall	0.5893	0.4942	0.5669
v3a3	Overall	0.4018	0.4925	0.367
v3a9	Overall	0.3571	0.4813	0.4958
v3a17	Overall	0.6964	0.4619	0.6245
v3b4	Overall	0.625	0.4863	0.6604
v3b12	Overall	0.6607	0.4756	0.469
v3b15	Overall	0.2679	0.4448	-0.1242
v3b16	Overall	0.4375	0.4983	0.4936
v5a7	Overall	0.4911	0.5022	0.7128
v5a10	Overall	0.4375	0.4983	0.5347
v5a11	Overall	0.4018	0.4925	0.3999
v5a14	Overall	0.6518	0.4785	0.6391
v5a18	Overall	0.4375	0.4983	0.4404
v5b2	Overall	0.4375	0.4983	0.5553
v5b5	Overall	0.5625	0.4983	0.4242
v5b6	Overall	0.4732	0.5015	0.649
v5b8	Overall	0.5893	0.4942	0.7366
v5b13	Overall	0.6518	0.4785	0.8366

```
====================================================================
   TEST LEVEL STATISTICS – National Sample
====================================================================
Number of Items = 36
Number of Examinees = 112
Min = 40.0000
Max = 32.0000
Mean = 20.1429
Median = 20.0000
Standard Deviation = 5.9249
Interquartile Range = 8.0000
Skewness = 0-0.0293
Kurtosis = -0.4396
KR21 = 0.7686
=====================================================================
```

RELIABILITY ANALYSIS – National Sample

```
======================================================================
```

Method	**Estimate**	**95% Conf. Int.**		**SEM**
Guttman's L2	0.8101	(0.7560,	0.8572)	2.5932
Coefficient Alpha	0.7899	(0.7300,	0.8420)	2.728
Feldt-Gilmer	0.8024	(0.7460,	0.8514)	2.6457
Feldt-Brennan	0.7973	(0.7395,	0.8476)	2.6795
Raju's Beta	0.7899	(0.7300,	0.8420)	2.728

```
           RELIABILITY IF ITEM DELETED - National Sample
=================================================================
```

Item	L2	Alpha	F-G	F-B	Raju
m12	0.8198	0.8019	0.8124	0.8083	0.8019
m3	0.8085	0.7878	0.8012	0.7957	0.7878
m1	0.8099	0.7892	0.8028	0.7972	0.7892
m13	0.8135	0.7936	0.8058	0.801	0.7936
m18	0.8134	0.7936	0.8062	0.8011	0.7936
m9	0.8084	0.7875	0.8014	0.7957	0.7875
m4	0.8065	0.7858	0.799	0.7935	0.7858
m16	0.8086	0.7881	0.8015	0.796	0.7881
m14	0.8104	0.7906	0.8027	0.7977	0.7906
m6	0.8114	0.7911	0.804	0.7988	0.7911
m11	0.8164	0.7978	0.8094	0.8047	0.7978
m7	0.8061	0.7852	0.7983	0.7929	0.7852
m5	0.8136	0.794	0.8067	0.8016	0.794
m2	0.8131	0.7931	0.8056	0.8006	0.7931
m10	0.812	0.7918	0.8043	0.7993	0.7918
m15	0.8066	0.7853	0.7985	0.793	0.7853
m8	0.8076	0.7869	0.7998	0.7945	0.7869
m17	0.8083	0.7874	0.8007	0.7952	0.7874
v3a1	0.7994	0.7779	0.791	0.7854	0.7779
v3a3	0.8056	0.7845	0.7976	0.7922	0.7845
v3a9	0.8023	0.7806	0.7941	0.7884	0.7806
v3a17	0.7989	0.7774	0.7903	0.7848	0.7774
v3b4	0.7967	0.7752	0.788	0.7825	0.7752
v3b12	0.8024	0.7816	0.7942	0.7889	0.7816
v3b15	0.8169	0.7982	0.8092	0.8049	0.7982
v3b16	0.8013	0.7802	0.793	0.7876	0.7802
v5a7	0.7946	0.7726	0.7856	0.7799	0.7726
v5a10	0.8006	0.7789	0.7921	0.7864	0.7789
v5a11	0.8045	0.7834	0.7963	0.7909	0.7834
v5a14	0.7975	0.7762	0.7891	0.7836	0.7762
v5a18	0.8032	0.782	0.7953	0.7897	0.782
v5b2	0.7998	0.7782	0.7916	0.7859	0.7782
v5b5	0.8035	0.7825	0.7956	0.7901	0.7825
v5b6	0.7966	0.7749	0.7878	0.7822	0.7749
v5b8	0.7938	0.7723	0.785	0.7795	0.7723
v5b13	0.7916	0.7699	0.7823	0.777	0.7699

```
=================================================================
L2: Guttman's lambda-2
Alpha: Coefficient alpha
F-G: Feldt-Gilmer coeff.
F-B: Feldt-Brennan coef.
Raju: Raju's beta coeff.
```

DIF Analysis

Item	Chi-square	p-value	Valid N	E.S.	(95% CI)	Class
m12	9.61	0	118	3.39	(5.64, 1.13)	C+
m3	5.62	0.02	106	-2.74	(-0.44, -5.05)	B-
m1	2.3	0.13	119	-1.49	(0.52, -3.50)	A
m13	6.79	0.01	105	-3.02	(-0.70, -5.34)	B-
m18	1.68	0.2	126	1.29	(3.24, -0.66)	A
m9	1.54	0.22	119	1.17	(3.08, -0.73)	A
m4	9.58	0	98	-3.08	(-0.91 -5.25)	B-
m16	0.21	0.65	116	0.51	(2.60, -1.59)	A
m14	0.77	0.38	95	-1.15	(1.48, -3.78)	A
m6	0.24	0.62	128	-0.48	(1.39, -2.35)	A
m11	0.02	0.88	113	-0.17	(1.96, -2.29)	A
m7	9.2	0	99	-4.24	(-1.44, -7.04)	C-
m5	7.08	0.01	125	-2.97	(-0.71, -5.23)	B-
m2	0.05	0.82	126	-0.21	(1.62, -2.04)	A
m10	1.74	0.19	121	-1.4	(0.64, -3.45)	A
m15	0.18	0.67	116	0.43	(2.44, -1.58)	A
m8	0.11	0.74	104	0.38	(2.58, -1.81)	A
m17	0.03	0.85	122	-0.18	(1.72, -2.07)	A
v3a1	1.44	0.23	102	1.36	(3.60, -0.87)	A
v3a3	0.13	0.71	119	0.36	(2.32, -1.60)	A
v3a9	0.58	0.45	116	0.7	(2.60, -1.19)	A
v3a17	0.01	0.94	85	-0.08	(2.15, -2.31)	A
v3b4	0.63	0.43	107	-0.89	(1.30, -3.08)	A
v3b12	7.93	0	110	3.69	(6.82, 0.56)	B+
v3b15	8.11	0	117	2.32	(4.19, 0.46)	B+
v3b16	0.18	0.67	109	-0.51	(1.75, -2.77)	A
v5a7	0.49	0.48	92	-0.74	(1.47, -2.95)	A
v5a10	0.44	0.51	99	0.77	(2.98, -1.45)	A
v5a11	0	0.97	116	0.05	(2.13, -2.04)	A
v5a14	0.69	0.4	111	0.81	(2.89, -1.26)	A
v5a18	0.92	0.34	115	0.93	(2.84, -0.99)	A
v5b2	2.62	0.11	105	1.67	(3.73, -0.39)	A
v5b5	2.51	0.11	106	-1.83	(0.36, -4.02)	A
v5b6	0.99	0.32	110	1.08	(3.26, -1.10)	A
v5b8	0.93	0.33	106	1.04	(3.26, -1.19)	A
v5b13	0.59	0.44	90	0.95	(3.42, -1.52)	A

```
               Options
----------------------------------
Matching Variable: score DIF Group
Variable: dataset
Focal Group Code: 1.0 (VT)
Reference Group Code: 0.0 (NS)
```

Appendix W – IRTPRO output for Dimensionality Investigation by Sample

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: National Sample – all items; EFA1 (2PL)

Item	*Label*		*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e.*
1	m12	2	-0.62	0.25	1	-0.26	0.2	-0.42	0.35
2	m3	4	0.95	0.29	3	0.46	0.22	-0.48	0.25
3	m1	6	0.75	0.27	5	0.14	0.2	-0.18	0.27
4	m13	8	0.51	0.25	7	0.57	0.2	-1.11	0.63
5	m18	10	-0.1	0.21	9	-0.05	0.19	-0.5	2.14
6	m9	12	1.27	0.35	11	-0.43	0.23	0.34	0.19
7	m4	14	2.27	0.69	13	2.47	0.56	-1.09	0.19
8	m16	16	1.37	0.38	15	0.83	0.25	-0.61	0.19
9	m14	18	0.71	0.37	17	2.01	0.32	-2.86	1.34
10	m6	20	0.53	0.24	19	0.12	0.2	-0.23	0.38
11	m11	22	-0.09	0.26	21	1.3	0.23	13.81	37.38
12	m7	24	1.66	0.53	23	2.45	0.48	-1.48	0.32
13	m5	26	0.48	0.23	25	-0.05	0.2	0.1	0.41
14	m2	28	-0.03	0.21	27	-0.09	0.19	-3.37	28.11
15	m10	30	0.09	0.21	29	-0.07	0.19	0.84	2.93
16	m15	32	0.48	0.23	31	0.03	0.2	-0.06	0.4
17	m8	34	1.36	0.41	33	1.62	0.33	-1.19	0.29
18	m17	36	0.88	0.3	35	0.71	0.22	-0.81	0.32
19	v3a1	38	0.75	0.29	37	0.7	0.23	-0.94	0.4
20	v3a3	40	0.79	0.28	39	-0.24	0.21	0.31	0.29
21	v3a9	42	0.62	0.26	41	-0.42	0.21	0.68	0.44
22	v3a17	44	1.72	0.52	43	1.85	0.42	-1.07	0.23
23	v3b4	46	1.46	0.44	45	1.35	0.33	-0.92	0.24
24	v3b12	48	0.37	0.3	47	1.24	0.25	-3.31	2.58
25	v3b15	50	-0.97	0.34	49	-0.88	0.26	-0.91	0.34
26	v3b16	52	0.73	0.27	51	0.04	0.22	-0.05	0.3
27	v5a7	54	1.4	0.39	53	0.55	0.27	-0.39	0.19
28	v5a10	56	0.73	0.28	55	0.14	0.22	-0.19	0.31
29	v5a11	58	0.4	0.25	57	-0.05	0.21	0.13	0.54
30	v5a14	60	0.7	0.34	59	1.47	0.29	-2.09	0.94
31	v5a18	62	0.32	0.24	61	0.13	0.21	-0.4	0.72
32	v5b2	64	0.51	0.27	63	0.24	0.22	-0.48	0.48
33	v5b5	66	0.11	0.27	65	0.97	0.24	-9.06	23.11
34	v5b6	68	1.17	0.36	67	0.55	0.26	-0.47	0.23
35	v5b8	70	1.6	0.51	69	1.6	0.39	-1	0.24
36	v5b13	72	2.01	0.7	71	2.53	0.62	-1.26	0.25

Oblique CF-Quartimax Rotated Loadings for Group 1

Item	**Label**	***λ*1**	***s.e.***
1	m12	0.34	0.21
2	m3	-0.49	0.19
3	m1	-0.4	0.2
4	m13	-0.29	0.22
5	m18	0.06	0.21
6	m9	-0.6	0.18
7	m4	-0.8	0.15
8	m16	-0.63	0.18
9	m14	-0.38	0.29
10	m6	-0.3	0.21
11	m11	0.06	0.26
12	m7	-0.7	0.2
13	m5	-0.27	0.21
14	m2	0.02	0.21
15	m10	-0.05	0.21
16	m15	-0.27	0.21
17	m8	-0.62	0.2
18	m17	-0.46	0.21
19	v3a1	-0.4	0.22
20	v3a3	-0.42	0.21
21	v3a9	-0.34	0.22
22	v3a17	-0.71	0.18
23	v3b4	-0.65	0.19
24	v3b12	-0.22	0.28
25	v3b15	0.49	0.22
26	v3b16	-0.39	0.21
27	v5a7	-0.64	0.18
28	v5a10	-0.4	0.22
29	v5a11	-0.23	0.23
30	v5a14	-0.38	0.27
31	v5a18	-0.19	0.23
32	v5b2	-0.29	0.23
33	v5b5	-0.06	0.27
34	v5b6	-0.57	0.2
35	v5b8	-0.69	0.19
36	v5b13	-0.76	0.19

Likelihood-based Values and Goodness of Fit Statistics	
Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	4337.21 ± 0.64
Akaike Information Criterion (AIC):	4481.21 ± 0.64
Bayesian Information Criterion (BIC):	4676.94 ± 0.64

Summary of the Data and Control Parameters	
Sample Size	112
Number of Items	36
Number of Dimensions	1

Parameter Estimation Control Values	
Metropolis-Hastings Robbins-Monro Algorithm	
Maximum number of cycles:	2000
Convergence criterion:	1.00E-03
Convergence monitor window size:	3
Number of imputations:	1
Thinning:	0
Burn-in:	10
Number of initialization cycles:	200
Gain constant for initialization cycles:	0.1
Control parameter alpha for gain sequence:	1
Control parameter epsilon for gain sequence:	1
Metropolis sampler type:	Spherical RWM
Metropolis proposal density standard deviation:	1
Standard error computation algorithm:	Accumulation

Miscellaneous Control Values	
Print parameter numbers?	Yes
Z tolerance, max. abs. logit value:	50
Random number seed:	5036
Sample size for Monte Carlo likelihood computation:	10000
Number of processor cores used:	1
Number of cycles completed:	540
Maximum parameter change:	7.76E-04
Final acceptance rate:	0.45
Number of free parameters:	72

Processing times (in seconds)	
Optimization and standard error:	21.78
Log-likelihood simulations:	2.06
Total:	23.84

Convergence and Numerical Stability	
Engine status:	Normal termination
First-order test:	Convergence criteria satisfied
Condition number of information matrix:	2.26E+01
Second-order test:	Solution is a possible local maximum

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: VT Sample - all items; EFA1 (2PL)

Item	*Label*		*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e.*
1	m12	2	-1.04	0.64	1	1.65	0.54	1.59	0.83
2	m3	4	1.89	0.86	3	-1.03	0.52	0.54	0.27
3	m1	6	0.48	0.44	5	-0.68	0.37	1.41	1.38
4	m13	8	0.27	0.4	7	-0.53	0.35	2	3.2
5	m18	10	0.33	0.39	9	0.42	0.35	-1.28	1.8
6	m9	12	0.15	0.38	11	0.17	0.34	-1.13	3.48
7	m4	14	1.07	0.53	13	0.07	0.38	-0.06	0.36
8	m16	16	2.03	0.94	15	0.66	0.48	-0.33	0.25
9	m14	18	-0.26	0.51	17	1.59	0.46	6.06	11.46
10	m6	20	0.45	0.41	19	-0.18	0.35	0.39	0.83
11	m11	22	0.92	0.52	21	1.24	0.45	-1.34	0.73
12	m7	24	0.62	0.43	23	0.44	0.36	-0.71	0.72
13	m5	26	0.22	0.47	25	-1.18	0.41	5.4	11.5
14	m2	28	-0.32	0.39	27	-0.24	0.35	-0.76	1.38
15	m10	30	0.32	0.43	29	-0.8	0.37	2.47	3.35
16	m15	32	1.14	0.55	31	0.43	0.4	-0.38	0.38
17	m8	34	0.31	0.44	33	1.09	0.4	-3.51	4.91
18	m17	36	0.69	0.44	35	0.19	0.36	-0.28	0.54
19	v3a1	38	0.64	0.45	37	0.71	0.38	-1.11	0.88
20	v3a3	40	0.91	0.52	39	-0.61	0.39	0.68	0.51
21	v3a9	42	-0.28	0.39	41	-0.42	0.35	-1.49	2.34
22	v3a17	44	0.93	0.51	43	0.77	0.4	-0.82	0.54
23	v3b4	46	1.07	0.53	45	0.21	0.38	-0.19	0.37
24	v3b12	48	-1.21	1.08	47	3.35	1.18	2.77	1.85
25	v3b15	50	1.35	0.62	49	0.22	0.4	-0.16	0.3
26	v3b16	52	1.69	0.74	51	-0.79	0.47	0.47	0.27
27	v5a7	54	0.79	0.47	53	-0.46	0.37	0.58	0.54
28	v5a10	56	1.06	0.54	55	-0.07	0.38	0.07	0.36
29	v5a11	58	0.83	0.48	57	-0.33	0.37	0.4	0.48
30	v5a14	60	-0.33	0.46	59	1.24	0.41	3.71	5.06
31	v5a18	62	0.58	0.42	61	0.06	0.35	-0.11	0.62
32	v5b2	64	0.53	0.42	63	0.31	0.36	-0.59	0.79
33	v5b5	66	2.47	1.11	65	-0.56	0.52	0.23	0.2
34	v5b6	68	0.63	0.43	67	0.19	0.36	-0.3	0.6
35	v5b8	70	0.71	0.46	69	0.72	0.39	-1.03	0.77
36	v5b13	72	1.49	0.71	71	1.26	0.52	-0.85	0.39

Oblique CF-Quartimax Rotated Loadings for Group 1

Item	**Label**	***λ*1**	***s.e.***
1	m12	-0.52	0.4
2	m3	0.74	0.26
3	m1	0.27	0.39
4	m13	0.15	0.39
5	m18	0.19	0.37
6	m9	0.09	0.37
7	m4	0.53	0.32
8	m16	0.77	0.25
9	m14	-0.15	0.49
10	m6	0.26	0.37
11	m11	0.48	0.35
12	m7	0.34	0.36
13	m5	0.13	0.45
14	m2	-0.19	0.37
15	m10	0.19	0.4
16	m15	0.56	0.31
17	m8	0.18	0.42
18	m17	0.38	0.35
19	v3a1	0.35	0.37
20	v3a3	0.47	0.36
21	v3a9	-0.16	0.37
22	v3a17	0.48	0.34
23	v3b4	0.53	0.32
24	v3b12	-0.58	0.58
25	v3b15	0.62	0.3
26	v3b16	0.7	0.27
27	v5a7	0.42	0.35
28	v5a10	0.53	0.33
29	v5a11	0.44	0.35
30	v5a14	-0.19	0.44
31	v5a18	0.32	0.36
32	v5b2	0.3	0.37
33	v5b5	0.82	0.2
34	v5b6	0.35	0.35
35	v5b8	0.38	0.36
36	v5b13	0.66	0.3

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	1489.7 ± 0.38
Akaike Information Criterion (AIC):	1633.70 ± 0.38
Bayesian Information Criterion (BIC):	1745.69 ± 0.38

Summary of the Data and Control Parameters

Sample Size	35
Number of Items	36
Number of Dimensions	1

Parameter Estimation Control Values

Metropolis-Hastings Robbins-Monro Algorithm	
Maximum number of cycles:	2000
Convergence criterion:	1.00E-03
Convergence monitor window size:	3
Number of imputations:	1
Thinning:	0
Burn-in:	10
Number of initialization cycles:	200
Gain constant for initialization cycles:	0.1
Control parameter alpha for gain sequence:	1
Control parameter epsilon for gain sequence:	1
Metropolis sampler type:	Spherical RWM
Metropolis proposal density standard deviation:	1
Standard error computation algorithm:	Accumulation

Miscellaneous Control Values

Print parameter numbers?	Yes
Z tolerance, max. abs. logit value:	50
Random number seed:	5036
Sample size for Monte Carlo likelihood computation:	10000
Number of processor cores used:	1
Number of cycles completed:	425
Maximum parameter change:	-8.62E-04
Final acceptance rate:	0.39
Number of free parameters:	72

Processing times (in seconds)

Optimization and standard error:	12.9
Log-likelihood simulations:	0.86
Total:	13.76

Convergence and Numerical Stability

Engine status:	Normal termination
First-order test:	Convergence criteria satisfied
Condition number of information matrix:	1.95E+01
Second-order test:	Solution is a possible local maximum

IRTPRO Version 4.2

Output generated by IRTPRO estimation engine Version 5.20 (64-bit)

Project: National Sample - all items; EFA2 (2PL)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*c*	*s.e.*
1	m12	2	-0.91	0.3		-0.08	-----	1	-0.3	0.21
2	m3	4	0.51	0.26	5	0.87	0.31	3	0.49	0.22
3	m1	7	-0.02	0.28	8	1.49	0.46	6	0.2	0.24
4	m13	10	0.47	0.26	11	0.27	0.25	9	0.59	0.21
5	m18	13	-0.53	0.26	14	0.37	0.25	12	-0.06	0.2
6	m9	16	0.54	0.29	17	1.56	0.47	15	-0.44	0.25
7	m4	19	1.92	0.57	20	1.23	0.48	18	2.55	0.56
8	m16	22	1.85	0.52	23	0.42	0.31	21	1.03	0.31
9	m14	25	3.45	1.79	26	-1.27	0.87	24	4.49	1.9
10	m6	28	0.72	0.28	29	0.05	0.24	27	0.15	0.2
11	m11	31	-0.38	0.29	32	0.29	0.28	30	1.37	0.25
12	m7	34	0.95	0.45	35	1.6	0.59	33	2.61	0.55
13	m5	37	0.01	0.24	38	0.74	0.28	36	-0.04	0.2
14	m2	40	-0.3	0.24	41	0.27	0.24	39	-0.09	0.2
15	m10	43	-0.64	0.3	44	0.82	0.32	42	-0.08	0.22
16	m15	46	0.21	0.23	47	0.49	0.25	45	0.04	0.2
17	m8	49	1.38	0.43	50	0.5	0.33	48	1.72	0.35
18	m17	52	0.86	0.3	53	0.4	0.27	51	0.75	0.23
19	v3a1	55	0.37	0.26	56	0.67	0.3	54	0.73	0.23
20	v3a3	58	0.61	0.27	59	0.5	0.28	57	-0.22	0.22
21	v3a9	61	0.04	0.26	62	0.96	0.34	60	-0.42	0.23
22	v3a17	64	1.33	0.43	65	0.99	0.41	63	1.86	0.41
23	v3b4	67	1.39	0.42	68	0.58	0.35	66	1.41	0.34
24	v3b12	70	0.31	0.29	71	0.26	0.31	69	1.26	0.25
25	v3b15	73	-0.54	0.3	74	-0.85	0.35	72	-0.91	0.26
26	v3b16	76	0.78	0.29	77	0.25	0.26	75	0.05	0.22
27	v5a7	79	0.87	0.33	80	1.25	0.43	78	0.62	0.28
28	v5a10	82	0.2	0.25	83	0.99	0.35	81	0.19	0.24
29	v5a11	85	0.43	0.26	86	0.14	0.25	84	-0.04	0.21
30	v5a14	88	0.74	0.35	89	0.27	0.34	87	1.52	0.31
31	v5a18	91	0.05	0.24	92	0.4	0.27	90	0.15	0.21
32	v5b2	94	0.16	0.25	95	0.61	0.29	93	0.27	0.23
33	v5b5	97	-0.33	0.3	98	0.52	0.32	96	1.06	0.27
34	v5b6	100	1.08	0.35	101	0.58	0.32	99	0.59	0.27
35	v5b8	103	1.11	0.4	104	1.06	0.43	102	1.62	0.39
36	v5b13	106	1.23	0.58	107	2.38	0.9	105	3.06	0.86

Oblique CF-Quartimax Rotated Loadings for Group 1

Item	Label	*λ1*	*s.e.*	*λ2*	s.e.
1	m12	-0.18	0.06	-0.4	0.19
2	m3	0.51	0.21	0	0.23
3	m1	0.65	0.19	-0.35	0.22
4	m13	0.23	0.23	0.16	0.25
5	m18	0.11	0.22	-0.37	0.24
6	m9	0.72	0.18	-0.14	0.22
7	m4	0.63	0.19	0.38	0.22
8	m16	0.38	0.19	0.56	0.2
9	m14	-0.06	0.26	0.92	0.15
10	m6	0.14	0.22	0.33	0.24
11	m11	0.1	0.26	-0.28	0.28
12	m7	0.74	0.21	0.01	0.29
13	m5	0.39	0.22	-0.2	0.23
14	m2	0.1	0.22	-0.24	0.24
15	m10	0.31	0.22	-0.5	0.23
16	m15	0.31	0.22	-0.04	0.23
17	m8	0.4	0.23	0.43	0.24
18	m17	0.33	0.22	0.29	0.24
19	v3a1	0.41	0.23	-0.01	0.25
20	v3a3	0.36	0.23	0.15	0.25
21	v3a9	0.49	0.23	-0.24	0.25
22	v3a17	0.58	0.22	0.28	0.24
23	v3b4	0.43	0.23	0.41	0.24
24	v3b12	0.2	0.29	0.08	0.3
25	v3b15	-0.5	0.24	-0.02	0.26
26	v3b16	0.25	0.23	0.3	0.24
27	v5a7	0.65	0.2	0.05	0.23
28	v5a10	0.52	0.23	-0.17	0.23
29	v5a11	0.15	0.24	0.17	0.25
30	v5a14	0.26	0.3	0.27	0.3
31	v5a18	0.23	0.25	-0.09	0.25
32	v5b2	0.36	0.24	-0.1	0.25
33	v5b5	0.23	0.28	-0.31	0.28
34	v5b6	0.43	0.23	0.31	0.24
35	v5b8	0.6	0.23	0.19	0.25
36	v5b13	0.85	0.16	-0.04	0.24

Factor Correlation Matrix

	λ_1	λ_2
λ_1	1.00	0.23
λ_2	0.23	1.00

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	4256.94 ± 1.26
Akaike Information Criterion (AIC):	4470.94 ± 1.26
Bayesian Information Criterion (BIC):	4761.82 ± 1.26

Summary of the Data and Control Parameters

Sample Size	112
Number of Items	36
Number of Dimensions	2

Parameter Estimation Control Values

Metropolis-Hastings Robbins-Monro Algorithm	
Maximum number of cycles:	2000
Convergence criterion:	1.00E-03
Convergence monitor window size:	3
Number of imputations:	1
Thinning:	0
Burn-in:	10
Number of initialization cycles:	200
Gain constant for initialization cycles:	0.1
Control parameter alpha for gain sequence:	1
Control parameter epsilon for gain sequence:	1
Metropolis sampler type:	Spherical RWM
Metropolis proposal density standard deviation:	1
Standard error computation algorithm:	Accumulation

Miscellaneous Control Values

Print parameter numbers?	Yes
Z tolerance, max. abs. logit value:	50
Random number seed:	5036
Sample size for Monte Carlo likelihood computation:	10000
Number of processor cores used:	1
Number of cycles completed:	869
Maximum parameter change:	8.03E-04
Final acceptance rate:	0.28
Number of free parameters:	107

Processing times (in seconds)

Optimization and standard error:	31.8
Log-likelihood simulations:	2.36
Total:	34.16

Convergence and Numerical Stability

Engine status:	Normal termination
First-order test:	Convergence criteria satisfied
Condition number of information matrix:	1.86E+02
Second-order test:	Solution is a possible local maximum

IRTPRO Version 4.2

Output generated by IRTPRO estimation engine Version 5.20 (64-bit)

Project: VT Sample - all items; EFA2 (2PL)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*c*	*s.e.*
1	m12	2	0.57	0.54		-0.98	-----	1	1.72	0.5
2	m3	4	-0.08	0.5	5	1.83	0.8	3	-0.99	0.51
3	m1	7	0.36	0.43	8	0.64	0.45	6	-0.69	0.38
4	m13	10	0.92	0.52	11	0.37	0.45	9	-0.57	0.39
5	m18	13	-0.35	0.42	14	0.34	0.4	12	0.42	0.36
6	m9	16	0.18	0.38	17	0.16	0.37	15	0.19	0.34
7	m4	19	-0.48	0.44	20	0.82	0.47	18	0.05	0.37
8	m16	22	-0.35	0.58	23	2.34	1.11	21	0.69	0.51
9	m14	25	-0.41	0.51	26	-0.28	0.5	24	1.63	0.47
10	m6	28	-0.02	0.38	29	0.41	0.4	27	-0.17	0.35
11	m11	31	0.04	0.5	32	0.99	0.52	30	1.29	0.47
12	m7	34	0.81	0.49	35	0.63	0.46	33	0.56	0.4
13	m5	37	0.37	0.46	38	0.24	0.45	36	-1.19	0.42
14	m2	40	-0.23	0.41	41	-0.46	0.4	39	-0.28	0.36
15	m10	43	-0.46	0.43	44	0.37	0.43	42	-0.85	0.39
16	m15	46	-2.05	1.1	47	1.65	0.91	45	0.58	0.57
17	m8	49	-0.18	0.44	50	0.25	0.42	48	1.08	0.4
18	m17	52	26	48.69	53	17.46	36.63	51	6.78	9.67
19	v3a1	55	-0.4	0.44	56	0.55	0.42	54	0.71	0.38
20	v3a3	58	0.2	0.42	59	0.91	0.5	57	-0.59	0.39
21	v3a9	61	-0.33	0.41	62	-0.33	0.39	60	-0.45	0.36
22	v3a17	64	-1.08	0.59	65	1.04	0.55	63	0.88	0.46
23	v3b4	67	-0.06	0.41	68	0.91	0.48	66	0.21	0.37
24	v3b12	70	-2022.72	-----	71	-1579.45	-----	69	2760.4	-----
25	v3b15	73	-0.35	0.47	74	1.32	0.61	72	0.22	0.4
26	v3b16	76	-0.87	0.6	77	1.69	0.81	75	-0.89	0.52
27	v5a7	79	1.86	0.83	80	1.43	0.66	78	-0.54	0.49
28	v5a10	82	-0.99	0.58	83	1.21	0.6	81	-0.13	0.42
29	v5a11	85	0.07	0.43	86	1	0.52	84	-0.33	0.38
30	v5a14	88	2.3	1.27	89	-0.31	0.7	87	2.27	0.96
31	v5a18	91	0.5	0.44	92	0.76	0.46	90	0.11	0.37
32	v5b2	94	0.02	0.39	95	0.45	0.41	93	0.31	0.35
33	v5b5	97	0.28	0.59	98	2.65	1.32	96	-0.53	0.53
34	v5b6	100	0.41	0.43	101	0.7	0.46	99	0.24	0.37
35	v5b8	103	-0.67	0.53	104	0.8	0.48	102	0.78	0.42
36	v5b13	106	0.55	0.56	107	1.61	0.73	105	1.39	0.55

Oblique CF-Quartimax Rotated Loadings for Group 1

Item	**Label**	$\lambda 1$	***s.e.***	$\lambda 2$	**s.e.**
1	m12	0.01	0.39	0.56	0.19
2	m3	0.33	0.33	-0.62	0.28
3	m1	0.34	0.37	-0.17	0.38
4	m13	0.5	0.34	0.12	0.39
5	m18	-0.08	0.38	-0.27	0.38
6	m9	0.14	0.37	-0.01	0.37
7	m4	0	0.37	-0.49	0.33
8	m16	0.29	0.33	-0.73	0.25
9	m14	-0.28	0.45	-0.01	0.49
10	m6	0.11	0.37	-0.2	0.37
11	m11	0.27	0.42	-0.4	0.37
12	m7	0.52	0.33	-0.02	0.37
13	m5	0.25	0.41	0.01	0.45
14	m2	-0.24	0.38	0.13	0.37
15	m10	-0.12	0.39	-0.32	0.38
16	m15	-0.31	0.31	-0.82	0.21
17	m8	-0.02	0.42	-0.18	0.42
18	m17	1	-----	0.03	-----
19	v3a1	-0.04	0.39	-0.37	0.36
20	v3a3	0.32	0.36	-0.32	0.36
21	v3a9	-0.26	0.37	0.04	0.38
22	v3a17	-0.19	0.36	-0.65	0.28
23	v3b4	0.21	0.36	-0.4	0.34
24	v3b12	-1	-----	0.04	-----
25	v3b15	0.16	0.36	-0.59	0.3
26	v3b16	0.03	0.35	-0.74	0.26
27	v5a7	0.81	0.19	-0.02	0.34
28	v5a10	-0.12	0.36	-0.68	0.27
29	v5a11	0.28	0.35	-0.39	0.36
30	v5a14	0.65	0.36	0.56	0.33
31	v5a18	0.42	0.34	-0.17	0.36
32	v5b2	0.14	0.38	-0.2	0.37
33	v5b5	0.49	0.29	-0.63	0.24
34	v5b6	0.37	0.36	-0.18	0.37
35	v5b8	-0.1	0.4	-0.52	0.34
36	v5b13	0.53	0.33	-0.41	0.33

Factor Correlation Matrix

	λ_1	λ_2
λ_1	1.00	-0.11
λ_2	-0.11	1.00

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	2341.68 ± 0.74
Akaike Information Criterion (AIC):	2555.68 ± 0.74
Bayesian Information Criterion (BIC):	2722.10 ± 0.74

Summary of the Data and Control Parameters

Sample Size	35
Number of Items	36
Number of Dimensions	2

Parameter Estimation Control Values

Metropolis-Hastings Robbins-Monro Algorithm	
Maximum number of cycles:	2000
Convergence criterion:	1.00E-03
Convergence monitor window size:	3
Number of imputations:	1
Thinning:	0
Burn-in:	10
Number of initialization cycles:	200
Gain constant for initialization cycles:	0.1
Control parameter alpha for gain sequence:	1
Control parameter epsilon for gain sequence:	1
Metropolis sampler type:	Spherical RWM
Metropolis proposal density standard deviation:	1
Standard error computation algorithm:	Accumulation

Miscellaneous Control Values

Print parameter numbers?	Yes
Z tolerance, max. abs. logit value:	50
Random number seed:	5036
Sample size for Monte Carlo likelihood computation:	10000
Number of processor cores used:	1
Number of cycles completed:	2000
Maximum parameter change:	2.10E-01
Final acceptance rate:	0.20
Number of free parameters:	107

Processing times (in seconds)

Optimization and standard error:	40.81
Log-likelihood simulations:	0.97
Total:	41.78

Convergence and Numerical Stability

Engine status:	Normal termination
First-order test:	Convergence criteria satisfied
Condition number of information matrix:	-2.86E+01
Second-order test:	Solution is a possible local maximum

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: VT Sample - all items; EFA2 (2PL)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*a3*	*s.e.*		*c*	*s.e.*
1	m12	2	0.09	0.24		-0.08	-----		-0.38	-----	1	-0.25	0.19
2	m3	4	0.82	0.36	5	1.09	0.51		0	-----	3	0.51	0.25
3	m1	7	0.31	0.28	8	1.18	0.4	9	0.45	0.28	6	0.15	0.23
4	m13	11	0.41	0.26	12	0.26	0.25	13	0.29	0.24	10	0.59	0.21
5	m18	15	0.3	0.26	16	0.38	0.26	17	-0.47	0.25	14	-0.06	0.2
6	m9	19	0.84	0.39	20	0.81	0.39	21	0.88	0.31	18	-0.45	0.24
7	m4	23	0.9	0.39	24	0.37	0.4	25	1.93	0.59	22	2.49	0.56
8	m16	27	0.2	0.29	28	0.03	0.31	29	1.47	0.41	26	0.89	0.27
9	m14	31	0.47	0.42	32	-1.74	0.78	33	1.79	0.68	30	3.38	0.92
10	m6	35	0.03	0.24	36	-0.05	0.24	37	0.55	0.25	34	0.13	0.2
11	m11	39	-0.46	0.34	40	0.86	0.35	41	-0.37	0.3	38	1.53	0.3
12	m7	43	0.22	0.44	44	1.57	0.72	45	1.47	0.65	42	2.86	0.74
13	m5	47	0.26	0.25	48	0.46	0.26	49	0.25	0.23	46	-0.05	0.2
14	m2	51	-0.4	0.27	52	0.44	0.25	53	-0.06	0.23	50	-0.11	0.2
15	m10	55	0.11	0.25	56	0.64	0.27	57	-0.17	0.24	54	-0.09	0.2
16	m15	59	0.12	0.25	60	0.21	0.25	61	0.43	0.24	58	0.03	0.2
17	m8	63	1.05	0.44	64	-0.6	0.41	65	1.66	0.51	62	2.07	0.47
18	m17	67	-0.18	0.27	68	0.18	0.29	69	1.12	0.36	66	0.77	0.24
19	v3a1	71	-0.9	0.41	72	0.72	0.39	73	1.34	0.46	70	0.92	0.31
20	v3a3	75	0.31	0.27	76	0.52	0.31	77	0.47	0.26	74	-0.23	0.22
21	v3a9	79	0.23	0.28	80	0.86	0.34	81	0.39	0.27	78	-0.44	0.23
22	v3a17	83	1.67	0.59	84	-0.17	0.45	85	1.52	0.54	82	2.25	0.57
23	v3b4	87	-0.68	0.63	88	-0.33	0.56	89	3.4	1.74	86	2.17	0.92
24	v3b12	91	-0.07	0.29	92	0.28	0.31	93	0.36	0.3	90	1.26	0.25
25	v3b15	95	-0.34	0.28	96	-0.39	0.31	97	-0.81	0.32	94	-0.9	0.26
26	v3b16	99	0.02	0.26	100	0.08	0.27	101	0.85	0.3	98	0.04	0.22
27	v5a7	103	1.46	0.53	104	1.27	0.58	105	0.85	0.4	102	0.73	0.34
28	v5a10	107	0.68	0.31	108	0.57	0.32	109	0.43	0.27	106	0.19	0.23
29	v5a11	111	-0.14	0.26	112	0.33	0.26	113	0.36	0.26	110	-0.06	0.22
30	v5a14	115	-0.05	0.33	116	-0.28	0.38	117	1.16	0.43	114	1.64	0.35
31	v5a18	119	0.76	0.31	120	0.3	0.28	121	-0.11	0.27	118	0.19	0.23
32	v5b2	123	-2.02	0.75	124	1.64	0.6	125	1.42	0.69	122	0.39	0.39
33	v5b5	127	0.34	0.31	128	0.42	0.31	129	-0.21	0.31	126	1.06	0.27
34	v5b6	131	0.6	0.31	132	0.42	0.31	133	0.81	0.31	130	0.57	0.26
35	v5b8	135	15.2	-----	136	-5.53	-----	137	11.5	-----	134	13.79	-----
36	v5b13	139	0.71	0.49	140	1.75	0.75	141	2.15	0.82	138	3.24	0.93

Oblique CF-Quartimax Rotated Loadings for Group 1

Item	**Label**	$\lambda 1$	***s.e.***	$\lambda 2$	**s.e.**	$\lambda 3$	**s.e.**
1	m12	0.07	0.13	-0.2	0.17	-0.03	0.13
2	m3	0.1	0.19	-0.18	0.15	0.64	0.27
3	m1	0.14	0.21	0.14	0.23	0.58	0.22
4	m13	-0.13	0.24	-0.03	0.24	0.26	0.24
5	m18	0.19	0.23	-0.26	0.24	0.25	0.24
6	m9	-0.21	0.23	0.06	0.25	0.56	0.25
7	m4	-0.51	0.21	0.28	0.22	0.36	0.23
8	m16	-0.42	0.21	0.39	0.23	0.12	0.24
9	m14	-0.79	0.18	0.19	0.22	-0.36	0.25
10	m6	-0.21	0.23	0.2	0.24	0.02	0.23
11	m11	0.51	0.23	0.12	0.29	0.23	0.25
12	m7	0	0.29	0.42	0.28	0.59	0.22
13	m5	0	0.24	0.04	0.24	0.31	0.23
14	m2	0.3	0.23	0.18	0.25	0.09	0.23
15	m10	0.24	0.23	-0.05	0.24	0.33	0.23
16	m15	-0.11	0.23	0.14	0.26	0.17	0.23
17	m8	-0.73	0.19	0.11	0.24	0.08	0.22
18	m17	-0.21	0.24	0.46	0.23	0.08	0.24
19	v3a1	0.08	0.25	0.7	0.2	0.11	0.25
20	v3a3	-0.06	0.25	0.1	0.25	0.36	0.27
21	v3a9	0.09	0.24	0.13	0.25	0.46	0.25
22	v3a17	-0.68	0.22	-0.07	0.24	0.32	0.25
23	v3b4	-0.47	0.2	0.73	0.18	-0.09	0.23
24	v3b12	0.01	0.29	0.2	0.29	0.14	0.3
25	v3b15	0.22	0.25	-0.2	0.25	-0.31	0.26
26	v3b16	-0.24	0.23	0.31	0.24	0.08	0.25
27	v5a7	-0.19	0.21	-0.09	0.22	0.73	0.22
28	v5a10	-0.14	0.24	-0.05	0.25	0.46	0.25
29	v5a11	0.05	0.24	0.23	0.25	0.14	0.26
30	v5a14	-0.4	0.26	0.38	0.27	-0.08	0.32
31	v5a18	-0.09	0.24	-0.31	0.25	0.36	0.25
32	v5b2	0.39	0.22	0.81	0.17	0.13	0.23
33	v5b5	0.11	0.27	-0.18	0.31	0.3	0.29
34	v5b6	-0.27	0.23	0.1	0.25	0.38	0.25
35	v5b8	-0.94	0.07	-0.2	0.15	0.23	0.18
36	v5b13	-0.17	0.24	0.39	0.21	0.64	0.2

Factor Correlation Matrix

	λ_1	λ_2	λ_3
λ_1	1.00	-0.15	-0.20
λ_2	-0.15	1.00	0.19
λ_3	-0.20	0.19	1.00

Likelihood-based Values and Goodness of Fit Statistics	
Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	4202.71 ± 2.13
Akaike Information Criterion (AIC):	4484.71 ± 2.13
Bayesian Information Criterion (BIC):	4868.02 ± 2.13

Summary of the Data and Control Parameters	
Sample Size	112
Number of Items	36
Number of Dimensions	3

Miscellaneous Control Values	
Print parameter numbers?	Yes
Z tolerance, max. abs. logit value:	50
Random number seed:	5036
Sample size for Monte Carlo likelihood computation:	10000
Number of processor cores used:	1
Number of cycles completed:	483
Maximum parameter change:	-4.05E+04
Final acceptance rate:	0.18
Number of free parameters:	141

Processing times (in seconds)	
Optimization and standard error:	21.03
Log-likelihood simulations:	2.34
Total:	23.37

Convergence and Numerical Stability	
Engine status:	Normal termination
First-order test:	Convergence criteria satisfied
Condition number of information matrix:	-1.94E-03
Second-order test:	Solution is a possible local maximum

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: VT Sample - all items; EFA3 (2PL)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*a3*	*s.e.*		*c*	*s.e.*
1	m12	2	1.67	0.87		-1	-----		0.36	-----	1	2	0.66
2	m3	4	0.14	0.53	5	1.98	0.78		-0.89	-----	3	-0.97	0.5
3	m1	7	0.18	0.41	8	0.53	0.44	9	-0.36	0.43	6	-0.68	0.38
4	m13	11	0.56	0.44	12	0.28	0.4	13	-0.17	0.43	10	-0.59	0.38
5	m18	15	-0.85	0.48	16	0.28	0.41	17	0.34	0.45	14	0.56	0.4
6	m9	19	0.21	0.39	20	0.21	0.38	21	-0.41	0.43	18	0.19	0.35
7	m4	23	0.45	0.48	24	1.21	0.58	25	-0.26	0.45	22	0.15	0.4
8	m16	27	-0.33	0.54	28	1.75	0.77	29	0.13	0.47	26	0.77	0.48
9	m14	31	30.39	1.59	32	-0.96	3.2	33	58.66	1.88	30	48.18	3.78
10	m6	35	-0.1	0.39	36	0.4	0.4	37	-0.44	0.42	34	-0.14	0.36
11	m11	39	-0.38	0.51	40	0.84	0.5	41	-0.18	0.47	38	1.34	0.48
12	m7	43	25.18	-----	44	17.44	7.31	45	-9.12	12.99	42	7.36	2.13
13	m5	47	1.37	0.73	48	0.69	0.58	49	-0.62	0.61	46	-1.64	0.6
14	m2	51	0.01	0.39	52	-0.41	0.39	53	-0.29	0.41	50	-0.25	0.36
15	m10	55	-0.37	0.43	56	0.27	0.41	57	0.31	0.46	54	-0.79	0.38
16	m15	59	-123.6	9.94	60	89.76	6.84	61	35.83	4.98	58	42.14	8.75
17	m8	63	1.08	0.72	64	0.76	0.6	65	0.75	0.6	62	1.44	0.58
18	m17	67	0.66	0.47	68	0.76	0.49	69	-0.48	0.48	66	0.24	0.39
19	v3a1	71	0.47	0.51	72	0.95	0.52	73	0.69	0.52	70	0.89	0.45
20	v3a3	75	-2.87	3.95	76	6.11	6.65	77	-11.17	10.59	74	-3.4	3.57
21	v3a9	79	-1.15	0.7	80	-0.69	0.53	81	-0.7	0.55	78	-0.56	0.45
22	v3a17	83	-0.15	0.57	84	1.41	0.68	85	1.24	0.72	82	1.1	0.56
23	v3b4	87	0.2	0.49	88	1.3	0.64	89	-1.14	0.65	86	0.35	0.43
24	v3b12	91	7653.43	-----	92	-38825.8	-----	93	41949.86	-----	90	48271.19	-----
25	v3b15	95	0.33	0.51	96	1.45	0.66	97	0.36	0.49	94	0.34	0.43
26	v3b16	99	-0.37	0.46	100	1.45	0.63	101	0.17	0.49	98	-0.66	0.44
27	v5a7	103	1.05	0.61	104	1.01	0.56	105	-1.05	0.65	102	-0.55	0.44
28	v5a10	107	-1.46	0.78	108	1.33	0.73	109	1.26	0.87	106	0	0.5
29	v5a11	111	0.09	0.42	112	1.08	0.51	113	0.21	0.44	110	-0.3	0.39
30	v5a14	115	1.47	0.75	116	-0.37	0.58	117	-0.34	0.61	114	1.56	0.57
31	v5a18	119	0.68	0.54	120	0.95	0.52	121	0.98	0.59	118	0.11	0.42
32	v5b2	123	0.94	0.59	124	0.79	0.51	125	0.75	0.54	122	0.41	0.42
33	v5b5	127	-0.26	0.51	128	1.75	0.8	129	-0.19	0.48	126	-0.34	0.45
34	v5b6	131	-0.61	0.45	132	0.63	0.43	133	-0.39	0.45	130	0.3	0.38
35	v5b8	135	-0.3	0.46	136	0.68	0.45	137	0.69	0.52	134	0.83	0.42
36	v5b13	139	-0.03	0.51	140	1.33	0.64	141	-0.08	0.47	138	1.3	0.51

Oblique CF-Quartimax Rotated Loadings for Group 1

Item	Label	$\lambda 1$	*s.e.*	$\lambda 2$	s.e.	$\lambda 3$	s.e.
1	m12	-0.4	0.05	-0.6	0.14	0.22	0.35
2	m3	0.39	0.14	0.31	0.3	0.54	0.27
3	m1	0.19	0.38	0.05	0.38	0.29	0.37
4	m13	0.01	0.4	-0.13	0.38	0.34	0.37
5	m18	0.01	0.38	0.44	0.33	-0.26	0.35
6	m9	0.19	0.38	-0.1	0.38	0.19	0.37
7	m4	0.12	0.34	0.19	0.34	0.54	0.33
8	m16	0.09	0.32	0.58	0.29	0.35	0.37
9	m14	-0.99	0.01	0.09	0.05	0.24	0.07
10	m6	0.28	0.36	0.08	0.37	0.13	0.37
11	m11	0.21	0.4	0.37	0.39	0.14	0.44
12	m7	0.08	-----	-0.25	0.09	0.98	-----
13	m5	0.09	0.38	-0.28	0.38	0.66	0.31
14	m2	0.12	0.38	-0.23	0.36	-0.13	0.38
15	m10	-0.07	0.41	0.3	0.39	-0.08	0.4
16	m15	0.13	0.04	0.97	0.01	-0.26	0.05
17	m8	-0.42	0.37	0.07	0.36	0.54	0.39
18	m17	0.16	0.37	-0.05	0.35	0.51	0.34
19	v3a1	-0.32	0.36	0.3	0.34	0.42	0.37
20	v3a3	0.93	0.12	0.08	0.25	0.23	0.43
21	v3a9	0.42	0.34	-0.02	0.34	-0.55	0.36
22	v3a17	-0.36	0.35	0.62	0.28	0.26	0.39
23	v3b4	0.49	0.31	0.11	0.34	0.46	0.32
24	v3b12	-0.82	-----	-0.22	-----	-0.42	-----
25	v3b15	-0.11	0.36	0.4	0.31	0.5	0.35
26	v3b16	0.07	0.37	0.56	0.29	0.28	0.34
27	v5a7	0.31	0.34	-0.17	0.34	0.63	0.27
28	v5a10	-0.17	0.36	0.81	0.22	-0.14	0.31
29	v5a11	-0.04	0.36	0.37	0.34	0.37	0.34
30	v5a14	-0.1	0.44	-0.57	0.34	0.4	0.41
31	v5a18	-0.45	0.34	0.28	0.33	0.45	0.34
32	v5b2	-0.41	0.35	0.12	0.33	0.51	0.35
33	v5b5	0.21	0.33	0.51	0.32	0.39	0.32
34	v5b6	0.34	0.36	0.32	0.36	-0.01	0.36
35	v5b8	-0.23	0.39	0.47	0.34	0.07	0.39
36	v5b13	0.12	0.36	0.41	0.34	0.39	0.4

Factor Correlation Matrix

	λ_1	λ_2	λ_3
λ_1	1.00	0.10	0.07
λ_2	0.10	1.00	0.09
λ_3	0.07	0.09	1.00

Likelihood-based Values and Goodness of Fit Statistics	
Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	2287.34 ± 1.69
Akaike Information Criterion (AIC):	2569.34 ± 1.69
Bayesian Information Criterion (BIC):	2788.64 ± 1.69

Summary of the Data and Control Parameters	
Sample Size	35
Number of Items	36
Number of Dimensions	3

Miscellaneous Control Values	
Print parameter numbers?	Yes
Z tolerance, max. abs. logit value:	50
Random number seed:	5036
Sample size for Monte Carlo likelihood computation:	10000
Number of processor cores used:	1
Number of cycles completed:	636
Maximum parameter change:	8.51E-04
Final acceptance rate:	0.08
Number of free parameters:	141

Processing times (in seconds)	
Optimization and standard error:	17.21
Log-likelihood simulations:	0.89
Total:	18.1

Convergence and Numerical Stability	
Engine status:	Normal termination
First-order test:	Convergence criteria satisfied
Condition number of information matrix:	-1.92E+00
Second-order test:	Solution is a possible local maximum

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: National Sample - all items; CFA2 (2PL)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*c*	*s.e.*
1	m12	2	-0.59	0.27		0	-----	1	-0.27	0.2
2	m3	4	0.89	0.31		0	-----	3	0.46	0.22
3	m1	6	0.51	0.26		0	-----	5	0.14	0.2
4	m13	8	0.47	0.26		0	-----	7	0.57	0.2
5	m18	10	-0.25	0.24		0	-----	9	-0.05	0.19
6	m9	12	0.88	0.31		0	-----	11	-0.38	0.21
7	m4	14	2.04	0.71		0	-----	13	2.36	0.57
8	m16	16	1.97	0.6		0	-----	15	1.01	0.32
9	m14	18	1.13	0.45		0	-----	17	2.27	0.42
10	m6	20	0.76	0.31		0	-----	19	0.14	0.21
11	m11	22	0.13	0.28		0	-----	21	1.3	0.23
12	m7	24	1.45	0.49		0	-----	23	2.38	0.46
13	m5	26	0.37	0.24		0	-----	25	-0.04	0.19
14	m2	28	-0.23	0.24		0	-----	27	-0.09	0.19
15	m10	30	-0.25	0.24		0	-----	29	-0.07	0.19
16	m15	32	0.21	0.23		0	-----	31	0.04	0.19
17	m8	34	1.29	0.42		0	-----	33	1.62	0.33
18	m17	36	1.03	0.33		0	-----	35	0.76	0.24
19	v3a1		0	-----	38	0.68	0.3	37	0.67	0.23
20	v3a3		0	-----	40	0.76	0.3	39	-0.26	0.22
21	v3a9		0	-----	42	0.62	0.29	41	-0.44	0.22
22	v3a17		0	-----	44	1.46	0.53	43	1.67	0.4
23	v3b4		0	-----	46	1.34	0.46	45	1.28	0.32
24	v3b12		0	-----	48	0.45	0.32	47	1.25	0.26
25	v3b15		0	-----	50	-1.16	0.39	49	-0.9	0.27
26	v3b16		0	-----	52	0.95	0.34	51	0.02	0.23
27	v5a7		0	-----	54	1.5	0.47	53	0.5	0.28
28	v5a10		0	-----	56	0.6	0.29	55	0.11	0.22
29	v5a11		0	-----	58	0.53	0.28	57	-0.07	0.22
30	v5a14		0	-----	60	0.7	0.36	59	1.45	0.29
31	v5a18		0	-----	62	0.43	0.26	61	0.12	0.22
32	v5b2		0	-----	64	0.44	0.28	63	0.22	0.22
33	v5b5		0	-----	66	0.16	0.29	65	0.97	0.24
34	v5b6		0	-----	68	1.14	0.4	67	0.49	0.26
35	v5b8		0	-----	70	1.37	0.51	69	1.44	0.37
36	v5b13		0	-----	72	1.6	0.65	71	2.2	0.54

Oblique CF-Quartimax Rotated Loadings for Group 1

Item	**Label**	$\lambda 1$	***s.e.***	$\lambda 2$	**s.e.**
1	m12	-0.33	0.23	0	0
2	m3	0.47	0.22	0	0
3	m1	0.29	0.23	0	0
4	m13	0.26	0.24	0	0
5	m18	-0.14	0.23	0	0
6	m9	0.46	0.22	0	0
7	m4	0.77	0.19	0	0
8	m16	0.76	0.17	0	0
9	m14	0.55	0.26	0	0
10	m6	0.41	0.24	0	0
11	m11	0.08	0.27	0	0
12	m7	0.65	0.22	0	0
13	m5	0.21	0.22	0	0
14	m2	-0.13	0.23	0	0
15	m10	-0.15	0.24	0	0
16	m15	0.12	0.23	0	0
17	m8	0.6	0.21	0	0
18	m17	0.52	0.2	0	0
19	v3a1	0	0	0.37	0.24
20	v3a3	0	0	0.41	0.23
21	v3a9	0	0	0.34	0.24
22	v3a17	0	0	0.65	0.23
23	v3b4	0	0	0.62	0.22
24	v3b12	0	0	0.26	0.29
25	v3b15	0	0	-0.56	0.22
26	v3b16	0	0	0.49	0.22
27	v5a7	0	0	0.66	0.2
28	v5a10	0	0	0.33	0.25
29	v5a11	0	0	0.3	0.24
30	v5a14	0	0	0.38	0.28
31	v5a18	0	0	0.25	0.24
32	v5b2	0	0	0.25	0.25
33	v5b5	0	0	0.09	0.28
34	v5b6	0	0	0.56	0.23
35	v5b8	0	0	0.63	0.24
36	v5b13	0	0	0.68	0.25

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	4390.01 ± 0.98
Akaike Information Criterion (AIC):	4534.01 ± 0.98
Bayesian Information Criterion (BIC):	4729.74 ± 0.98

Summary of the Data and Control Parameters

Sample Size	112
Number of Items	36
Number of Dimensions	2

Miscellaneous Control Values

Print parameter numbers?	Yes
Z tolerance, max. abs. logit value:	50
Random number seed:	5036
Sample size for Monte Carlo likelihood computation:	10000
Number of processor cores used:	1
Number of cycles completed:	558
Maximum parameter change:	-1.09E-06
Final acceptance rate:	0.32
Number of free parameters:	72

Processing times (in seconds)

Optimization and standard error:	21.34
Log-likelihood simulations:	2.31
Total:	23.65

Convergence and Numerical Stability

Engine status:	Normal termination
First-order test:	Convergence criteria satisfied
Condition number of information matrix:	2.07E+01
Second-order test:	Solution is a possible local maximum

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: VT Sample - all items; CFA2 (2PL)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*c*	*s.e.*
1	m12	2	-0.32	0.49		0	-----	1	1.41	0.44
2	m3	4	24.4	57.29		0	-----	3	-9.32	19.78
3	m1	6	1.22	0.64		0	-----	5	-0.83	0.44
4	m13	8	0	0.4		0	-----	7	-0.53	0.35
5	m18	10	0.08	0.4		0	-----	9	0.41	0.35
6	m9	12	1.08	0.58		0	-----	11	0.21	0.39
7	m4	14	1.46	0.75		0	-----	13	0.08	0.41
8	m16	16	0.98	0.57		0	-----	15	0.55	0.4
9	m14	18	-0.31	0.51		0	-----	17	1.6	0.46
10	m6	20	0.53	0.45		0	-----	19	-0.18	0.35
11	m11	22	0.24	0.46		0	-----	21	1.08	0.39
12	m7	24	0.69	0.49		0	-----	23	0.45	0.37
13	m5	26	0.77	0.57		0	-----	25	-1.28	0.46
14	m2	28	-0.2	0.41		0	-----	27	-0.24	0.35
15	m10	30	0.48	0.46		0	-----	29	-0.82	0.38
16	m15	32	0.41	0.42		0	-----	31	0.37	0.36
17	m8	34	0.07	0.44		0	-----	33	1.06	0.39
18	m17	36	0.52	0.43		0	-----	35	0.19	0.35
19	v3a1		0	-----	38	0.77	0.49	37	0.73	0.39
20	v3a3		0	-----	40	0.55	0.45	39	-0.57	0.37
21	v3a9		0	-----	42	-0.29	0.41	41	-0.41	0.35
22	v3a17		0	-----	44	1.04	0.59	43	0.79	0.42
23	v3b4		0	-----	46	0.57	0.45	45	0.18	0.36
24	v3b12		0	-----	48	-0.59	0.93	47	2.95	0.87
25	v3b15		0	-----	50	1.24	0.63	49	0.21	0.4
26	v3b16		0	-----	52	1.2	0.63	51	-0.69	0.43
27	v5a7		0	-----	54	0.58	0.45	53	-0.44	0.37
28	v5a10		0	-----	56	1.23	0.62	55	-0.09	0.4
29	v5a11		0	-----	58	1.14	0.61	57	-0.38	0.4
30	v5a14		0	-----	60	-0.17	0.46	59	1.22	0.41
31	v5a18		0	-----	62	0.85	0.51	61	0.06	0.37
32	v5b2		0	-----	64	0.81	0.5	63	0.32	0.37
33	v5b5		0	-----	66	2.09	1.09	65	-0.52	0.52
34	v5b6		0	-----	68	0.75	0.48	67	0.19	0.37
35	v5b8		0	-----	70	0.72	0.49	69	0.72	0.39
36	v5b13		0	-----	72	1.78	0.97	71	1.39	0.63

Oblique CF-Quartimax Rotated Loadings for Group 1

Item	**Label**	$\lambda 1$	***s.e.***	$\lambda 2$	**s.e.**
1	m12	-0.19	0.47	0	0
2	m3	1	0.02	0	0
3	m1	0.58	0.34	0	0
4	m13	0	0.4	0	0
5	m18	0.05	0.4	0	0
6	m9	0.54	0.35	0	0
7	m4	0.65	0.33	0	0
8	m16	0.5	0.37	0	0
9	m14	-0.18	0.49	0	0
10	m6	0.3	0.39	0	0
11	m11	0.14	0.45	0	0
12	m7	0.38	0.39	0	0
13	m5	0.41	0.43	0	0
14	m2	-0.12	0.4	0	0
15	m10	0.27	0.41	0	0
16	m15	0.23	0.39	0	0
17	m8	0.04	0.44	0	0
18	m17	0.29	0.38	0	0
19	v3a1	0	0	0.41	0.37
20	v3a3	0	0	0.31	0.39
21	v3a9	0	0	-0.17	0.39
22	v3a17	0	0	0.52	0.37
23	v3b4	0	0	0.32	0.38
24	v3b12	0	0	-0.33	0.78
25	v3b15	0	0	0.59	0.33
26	v3b16	0	0	0.58	0.34
27	v5a7	0	0	0.33	0.38
28	v5a10	0	0	0.59	0.33
29	v5a11	0	0	0.56	0.35
30	v5a14	0	0	-0.1	0.46
31	v5a18	0	0	0.45	0.36
32	v5b2	0	0	0.43	0.36
33	v5b5	0	0	0.78	0.27
34	v5b6	0	0	0.4	0.37
35	v5b8	0	0	0.39	0.38
36	v5b13	0	0	0.72	0.32

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	1508.39 ± 0.66
Akaike Information Criterion (AIC):	1652.39 ± 0.66
Bayesian Information Criterion (BIC):	1746.37 ± 0.66

Summary of the Data and Control Parameters

Sample Size	35
Number of Items	36
Number of Dimensions	2

Miscellaneous Control Values

Print parameter numbers?	Yes
Z tolerance, max. abs. logit value:	50
Random number seed:	5036
Sample size for Monte Carlo likelihood computation:	10000
Number of processor cores used:	1
Number of cycles completed:	987
Maximum parameter change:	9.55E-04
Final acceptance rate:	0.23
Number of free parameters:	72

Processing times (in seconds)

Optimization and standard error:	28.44
Log-likelihood simulations:	0.94
Total:	29.38

Convergence and Numerical Stability

Engine status:	Normal termination
First-order test:	Convergence criteria satisfied
Condition number of information matrix:	3.08E+04
Second-order test:	Solution is a possible local maximum

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: National Sample - all items; Bifactor

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*a3*	*s.e.*		*c*	*s.e.*
1	m12	2	-0.56	0.24	3	-0.65	0.31		0	-----	1	-0.26	0.21
2	m3	5	0.92	0.29	6	-0.08	0.27		0	-----	4	0.44	0.22
3	m1	8	1.1	0.36	9	-0.97	0.37		0	-----	7	0.17	0.24
4	m13	11	0.49	0.24	12	0.32	0.29		0	-----	10	0.57	0.21
5	m18	14	-0.03	0.23	15	-0.74	0.31		0	-----	13	-0.06	0.21
6	m9	17	1.69	0.48	18	-0.77	0.38		0	-----	16	-0.53	0.27
7	m4	20	2.23	0.67	21	0.96	0.51		0	-----	19	2.57	0.59
8	m16	23	1.36	0.39	24	0.96	0.42		0	-----	22	0.87	0.28
9	m14	26	0.92	0.63	27	2.65	1.43		0	-----	25	3.56	1.31
10	m6	29	0.49	0.24	30	0.51	0.29		0	-----	28	0.12	0.2
11	m11	32	-0.07	0.26	33	-0.33	0.31		0	-----	31	1.34	0.24
12	m7	35	1.54	0.52	36	0.04	0.39		0	-----	34	2.34	0.45
13	m5	38	0.59	0.25	39	-0.42	0.29		0	-----	37	-0.06	0.2
14	m2	41	0.01	0.22	42	-0.58	0.29		0	-----	40	-0.09	0.2
15	m10	44	0.2	0.25	45	-0.91	0.36		0	-----	43	-0.08	0.22
16	m15	47	0.52	0.24	48	-0.34	0.26		0	-----	46	0.02	0.2
17	m8	50	1.32	0.4	51	0.66	0.37		0	-----	49	1.66	0.34
18	m17	53	0.82	0.29	54	0.5	0.3		0	-----	52	0.71	0.23
19	v3a1	56	0.86	0.35		0	-----	57	1.15	0.48	55	0.86	0.29
20	v3a3	59	0.77	0.28		0	-----	60	0.4	0.31	58	-0.25	0.22
21	v3a9	62	0.66	0.26		0	-----	63	-0.12	0.3	61	-0.43	0.22
22	v3a17	65	2.05	0.68		0	-----	66	-0.77	0.46	64	2.1	0.53
23	v3b4	68	2.12	1		0	-----	69	1.88	1.14	67	1.95	0.79
24	v3b12	71	0.37	0.29		0	-----	72	0.13	0.34	70	1.24	0.25
25	v3b15	74	-1.02	0.35		0	-----	75	-0.24	0.33	73	-0.88	0.26
26	v3b16	77	0.79	0.32		0	-----	78	0.84	0.41	76	0.03	0.24
27	v5a7	80	1.52	0.44		0	-----	81	-0.43	0.38	79	0.58	0.28
28	v5a10	83	0.84	0.31		0	-----	84	-0.47	0.34	82	0.16	0.23
29	v5a11	86	0.36	0.25		0	-----	87	0.43	0.33	85	-0.08	0.22
30	v5a14	89	0.7	0.35		0	-----	90	0.37	0.36	88	1.48	0.3
31	v5a18	92	0.41	0.27		0	-----	93	-0.64	0.36	91	0.17	0.23
32	v5b2	95	0.59	0.31		0	-----	96	1.16	0.46	94	0.25	0.26
33	v5b5	98	0.18	0.28		0	-----	99	-0.4	0.35	97	1.02	0.26
34	v5b6	101	1.1	0.34		0	-----	102	-0.08	0.31	100	0.53	0.25
35	v5b8	104	2.37	0.88		0	-----	105	-1.3	0.67	103	2.21	0.69
36	v5b13	107	2.13	0.74		0	-----	108	0.4	0.46	106	2.62	0.66

Oblique CF-Quartimax Rotated Loadings for Group 1

Item	**Label**	$\lambda 1$	***s.e.***	$\lambda 2$	**s.e.**	$\lambda 3$	**s.e.**
1	m12	-0.29	0.2	-0.34	0.24	0	0
2	m3	0.48	0.2	-0.04	0.24	0	0
3	m1	0.49	0.19	-0.43	0.21	0	0
4	m13	0.27	0.21	0.18	0.26	0	0
5	m18	-0.02	0.21	-0.4	0.24	0	0
6	m9	0.67	0.16	-0.3	0.21	0	0
7	m4	0.75	0.15	0.33	0.24	0	0
8	m16	0.57	0.17	0.4	0.23	0	0
9	m14	0.28	0.25	0.81	0.23	0	0
10	m6	0.27	0.2	0.28	0.25	0	0
11	m11	-0.04	0.25	-0.19	0.29	0	0
12	m7	0.67	0.21	0.02	0.29	0	0
13	m5	0.32	0.21	-0.23	0.24	0	0
14	m2	0	0.21	-0.32	0.24	0	0
15	m10	0.1	0.21	-0.47	0.25	0	0
16	m15	0.29	0.21	-0.19	0.23	0	0
17	m8	0.59	0.19	0.29	0.25	0	0
18	m17	0.42	0.21	0.25	0.24	0	0
19	v3a1	0.39	0.21	0	0	0.52	0.25
20	v3a3	0.4	0.21	0	0	0.21	0.26
21	v3a9	0.36	0.21	0	0	-0.06	0.27
22	v3a17	0.74	0.17	0	0	-0.28	0.24
23	v3b4	0.64	0.17	0	0	0.57	0.26
24	v3b12	0.21	0.27	0	0	0.07	0.33
25	v3b15	-0.51	0.22	0	0	-0.12	0.28
26	v3b16	0.38	0.21	0	0	0.41	0.28
27	v5a7	0.65	0.18	0	0	-0.19	0.26
28	v5a10	0.43	0.21	0	0	-0.24	0.28
29	v5a11	0.2	0.23	0	0	0.24	0.29
30	v5a14	0.37	0.27	0	0	0.2	0.31
31	v5a18	0.22	0.23	0	0	-0.34	0.29
32	v5b2	0.28	0.22	0	0	0.54	0.25
33	v5b5	0.1	0.26	0	0	-0.23	0.32
34	v5b6	0.54	0.2	0	0	-0.04	0.26
35	v5b8	0.74	0.17	0	0	-0.41	0.24
36	v5b13	0.77	0.18	0	0	0.15	0.27

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	4239.94 ± 1.69
Akaike Information Criterion (AIC):	4455.94 ± 1.69
Bayesian Information Criterion (BIC):	4749.54 ± 1.69

Summary of the Data and Control Parameters

Sample Size	112
Number of Items	36
Number of Dimensions	3

Miscellaneous Control Values

Print parameter numbers?	Yes
Z tolerance, max. abs. logit value:	50
Random number seed:	5036
Sample size for Monte Carlo likelihood computation:	10000
Number of processor cores used:	1
Number of cycles completed:	715
Maximum parameter change:	3.09E-04
Final acceptance rate:	0.45
Number of free parameters:	108

Processing times (in seconds)

Optimization and standard error:	38.42
Log-likelihood simulations:	2.18
Total:	40.61

Convergence and Numerical Stability

Engine status:	Normal termination
First-order test:	Convergence criteria satisfied
Condition number of information matrix:	9.53E+01
Second-order test:	Solution is a possible local maximum

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: VT Sample - all items; Bifactor

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*a3*	*s.e.*		*c*	*s.e.*
1	m12	2	-1.65	0.99	3	0.87	0.64		0	-----	1	2.12	0.82
2	m3	5	2.23	1.03	6	0.6	0.55		0	-----	4	-1.15	0.57
3	m1	8	0.49	0.44	9	0.16	0.42		0	-----	7	-0.69	0.37
4	m13	11	0.29	0.41	12	0.13	0.4		0	-----	10	-0.54	0.36
5	m18	14	0.4	0.47	15	-1.49	0.74		0	-----	13	0.66	0.47
6	m9	17	0.29	0.39	18	0.43	0.42		0	-----	16	0.18	0.35
7	m4	20	1.19	0.59	21	0.77	0.51		0	-----	19	0.07	0.41
8	m16	23	1.82	0.84	24	-0.37	0.52		0	-----	22	0.61	0.46
9	m14	26	-0.67	0.63	27	0.63	0.55		0	-----	25	1.81	0.57
10	m6	29	0.51	0.42	30	-0.02	0.4		0	-----	28	-0.18	0.35
11	m11	32	0.96	0.52	33	-0.4	0.53		0	-----	31	1.29	0.47
12	m7	35	3.33	3.42	36	7.19	7.44		0	-----	34	1.77	1.86
13	m5	38	0.76	0.85	39	8.27	-----		0	-----	37	-6	-----
14	m2	41	-0.32	0.39	42	0.26	0.41		0	-----	40	-0.26	0.36
15	m10	44	0.2	0.42	45	-0.46	0.43		0	-----	43	-0.81	0.38
16	m15	47	1.13	0.59	48	-0.97	0.59		0	-----	46	0.46	0.44
17	m8	50	0.22	0.47	51	0.89	0.53		0	-----	49	1.22	0.45
18	m17	53	0.77	0.46	54	-0.03	0.41		0	-----	52	0.19	0.36
19	v3a1	56	0.51	0.45		0	-----	57	0.94	0.57	55	0.78	0.42
20	v3a3	59	2.32	1.54		0	-----	60	-1.65	1.26	58	-1.01	0.7
21	v3a9	62	-0.24	0.39		0	-----	63	-0.28	0.45	61	-0.41	0.35
22	v3a17	65	0.75	0.5		0	-----	66	1.11	0.65	64	0.85	0.46
23	v3b4	68	1.68	0.83		0	-----	69	-0.79	0.66	67	0.24	0.44
24	v3b12	71	-25.23	2.41		0	-----	72	29.84	2.68	70	50.03	6.13
25	v3b15	74	1.24	0.63		0	-----	75	1.21	0.69	73	0.23	0.45
26	v3b16	77	1.48	0.7		0	-----	78	0.65	0.6	76	-0.8	0.47
27	v5a7	80	1.17	0.62		0	-----	81	-0.69	0.58	79	-0.53	0.42
28	v5a10	83	0.93	0.54		0	-----	84	1	0.63	82	-0.1	0.41
29	v5a11	86	0.75	0.48		0	-----	87	0.62	0.5	85	-0.36	0.39
30	v5a14	89	-0.31	0.47		0	-----	90	-0.12	0.51	88	1.25	0.41
31	v5a18	92	0.47	0.41		0	-----	93	0.5	0.47	91	0.05	0.36
32	v5b2	95	0.4	0.46		0	-----	96	1.24	0.66	94	0.35	0.42
33	v5b5	98	2.29	1.05		0	-----	99	0.34	0.58	97	-0.58	0.51
34	v5b6	101	0.81	0.49		0	-----	102	-0.23	0.46	100	0.2	0.37
35	v5b8	104	0.57	0.54		0	-----	105	1.53	0.89	103	0.93	0.53
36	v5b13	107	1.66	0.77		0	-----	108	0.02	0.59	106	1.29	0.53

Oblique CF-Quartimax Rotated Loadings for Group 1

Item	**Label**	***λ1***	***s.e.***	***λ2***	**s.e.**	***λ3***	**s.e.**
1	m12	-0.65	0.35	0.34	0.34	0	0
2	m3	0.78	0.24	0.21	0.31	0	0
3	m1	0.28	0.39	0.09	0.4	0	0
4	m13	0.17	0.39	0.08	0.39	0	0
5	m18	0.17	0.34	-0.65	0.31	0	0
6	m9	0.16	0.37	0.24	0.37	0	0
7	m4	0.54	0.31	0.35	0.33	0	0
8	m16	0.72	0.27	-0.15	0.34	0	0
9	m14	-0.34	0.49	0.32	0.43	0	0
10	m6	0.28	0.37	-0.01	0.38	0	0
11	m11	0.48	0.34	-0.2	0.44	0	0
12	m7	0.41	0.25	0.89	0.14	0	0
13	m5	0.09	0.33	0.98	-----	0	0
14	m2	-0.18	0.37	0.15	0.39	0	0
15	m10	0.11	0.4	-0.26	0.39	0	0
16	m15	0.5	0.32	-0.43	0.35	0	0
17	m8	0.11	0.41	0.46	0.37	0	0
18	m17	0.41	0.35	-0.02	0.37	0	0
19	v3a1	0.25	0.35	0	0	0.47	0.37
20	v3a3	0.7	0.27	0	0	-0.5	0.32
21	v3a9	-0.14	0.37	0	0	-0.16	0.43
22	v3a17	0.35	0.34	0	0	0.51	0.36
23	v3b4	0.67	0.27	0	0	-0.31	0.36
24	v3b12	-0.65	0.05	0	0	0.76	0.04
25	v3b15	0.51	0.31	0	0	0.5	0.34
26	v3b16	0.63	0.3	0	0	0.28	0.39
27	v5a7	0.54	0.32	0	0	-0.32	0.38
28	v5a10	0.43	0.33	0	0	0.46	0.38
29	v5a11	0.38	0.35	0	0	0.32	0.39
30	v5a14	-0.18	0.44	0	0	-0.07	0.5
31	v5a18	0.26	0.36	0	0	0.27	0.41
32	v5b2	0.18	0.35	0	0	0.58	0.35
33	v5b5	0.8	0.23	0	0	0.12	0.34
34	v5b6	0.43	0.35	0	0	-0.12	0.41
35	v5b8	0.24	0.35	0	0	0.65	0.36
36	v5b13	0.7	0.28	0	0	0.01	0.42

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	1426.16 ± 1.07
Akaike Information Criterion (AIC):	1642.16 ± 1.07
Bayesian Information Criterion (BIC):	1810.14 ± 1.07

Summary of the Data and Control Parameters

Sample Size	35
Number of Items	36
Number of Dimensions	3

Miscellaneous Control Values

Print parameter numbers?	Yes
Z tolerance, max. abs. logit value:	50
Random number seed:	5036
Sample size for Monte Carlo likelihood computation:	10000
Number of processor cores used:	1
Number of cycles completed:	1214
Maximum parameter change:	2.51E-04
Final acceptance rate:	0.35
Number of free parameters:	108

Processing times (in seconds)

Optimization and standard error:	42.92
Log-likelihood simulations:	0.98
Total:	43.9

Convergence and Numerical Stability

Engine status:	Normal termination
First-order test:	Convergence criteria satisfied
Condition number of information matrix:	-2.48E-01
Second-order test:	Solution is not a maximum; caution is advised

Appendix X – IRTPRO output for IRT Model Selection Investigation by Sample

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: National Sample - all items; Rasch

Item	*Label*	*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e.*
1	m12	1	-----	1	-0.28	0.21	0.28	0.21
2	m3	1	-----	2	0.45	0.22	-0.45	0.22
3	m1	1	-----	3	0.15	0.21	-0.15	0.21
4	m13	1	-----	4	0.61	0.22	-0.61	0.22
5	m18	1	-----	5	-0.06	0.21	0.06	0.21
6	m9	1	-----	6	-0.36	0.21	0.36	0.21
7	m4	1	-----	7	1.62	0.26	-1.62	0.26
8	m16	1	-----	8	0.7	0.22	-0.7	0.22
9	m14	1	-----	9	2.04	0.29	-2.04	0.29
10	m6	1	-----	10	0.14	0.21	-0.14	0.21
11	m11	1	-----	11	1.44	0.25	-1.44	0.25
12	m7	1	-----	12	1.95	0.29	-1.95	0.29
13	m5	1	-----	13	-0.04	0.21	0.04	0.21
14	m2	1	-----	14	-0.1	0.21	0.1	0.21
15	m10	1	-----	15	-0.08	0.21	0.08	0.21
16	m15	1	-----	16	0.04	0.21	-0.04	0.21
17	m8	1	-----	17	1.38	0.25	-1.38	0.25
18	m17	1	-----	18	0.7	0.22	-0.7	0.22
19	v3a1	1	-----	19	0.71	0.23	-0.71	0.23
20	v3a3	1	-----	20	-0.23	0.22	0.23	0.22
21	v3a9	1	-----	21	-0.42	0.23	0.42	0.23
22	v3a17	1	-----	22	1.38	0.26	-1.38	0.26
23	v3b4	1	-----	23	1.1	0.25	-1.1	0.25
24	v3b12	1	-----	24	1.34	0.26	-1.34	0.26
25	v3b15	1	-----	25	-0.85	0.24	0.85	0.24
26	v3b16	1	-----	26	0.05	0.23	-0.05	0.23
27	v5a7	1	-----	27	0.45	0.24	-0.45	0.24
28	v5a10	1	-----	28	0.15	0.23	-0.15	0.23
29	v5a11	1	-----	29	-0.04	0.23	0.04	0.23
30	v5a14	1	-----	30	1.49	0.28	-1.49	0.28
31	v5a18	1	-----	31	0.15	0.23	-0.15	0.23
32	v5b2	1	-----	32	0.26	0.24	-0.26	0.24
33	v5b5	1	-----	33	1.09	0.26	-1.09	0.26
34	v5b6	1	-----	34	0.47	0.24	-0.47	0.24
35	v5b8	1	-----	35	1.22	0.27	-1.22	0.27
36	v5b13	1	-----	36	1.74	0.3	-1.74	0.3

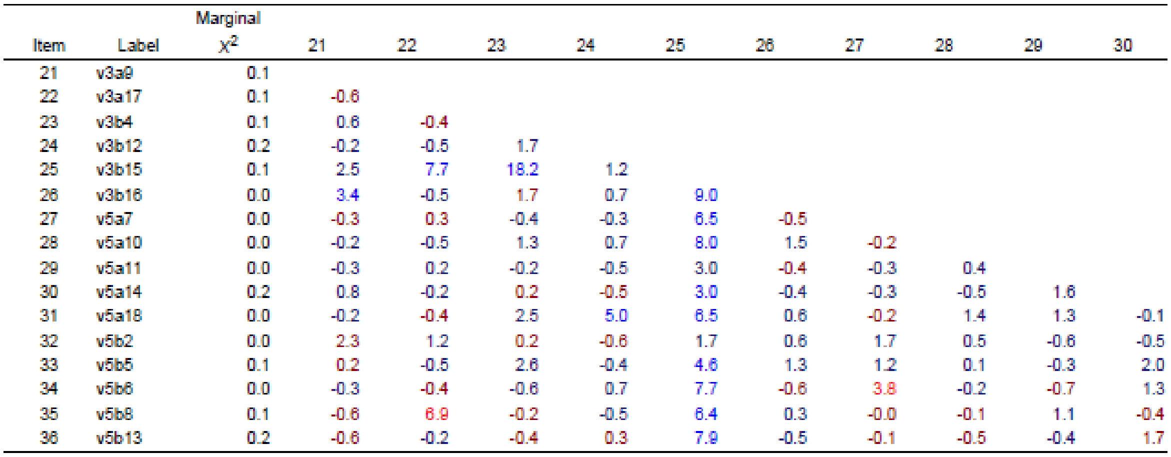
S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	*Probability*
1	m12	35.59	13	0.0007
2	m3	14.94	13	0.3129
3	m1	8.99	13	0.7746
4	m13	12.89	12	0.3788
5	m18	21.72	14	0.0843
6	m9	10.19	13	0.6794
7	m4	17.09	8	0.0291
8	m16	22.29	10	0.0136
9	m14	8.61	6	0.1965
10	m6	12.6	12	0.4006
11	m11	16.33	10	0.0904
12	m7	8.84	5	0.1153
13	m5	9.19	11	0.6051
14	m2	16.02	14	0.3111
15	m10	20.44	13	0.0846
16	m15	8.87	12	0.7148
17	m8	13.01	9	0.1617
18	m17	10.21	9	0.3353
19	v3a1	7.12	11	0.7902
20	v3a3	15.13	14	0.3714
21	v3a9	10.26	12	0.594
22	v3a17	15.1	9	0.088
23	v3b4	8.81	9	0.4568
24	v3b12	18.34	9	0.0314
25	v3b15	58.84	13	0.0001
26	v3b16	6.63	14	0.9483
27	v5a7	12.82	10	0.2332
28	v5a10	17.17	11	0.1025
29	v5a11	18.71	13	0.1319
30	v5a14	7.54	8	0.4805
31	v5a18	9.47	14	0.8003
32	v5b2	7.32	11	0.7732
33	v5b5	12.34	8	0.1363
34	v5b6	14.21	11	0.2211
35	v5b8	9.93	9	0.3576
36	v5b13	9.95	8	0.2707

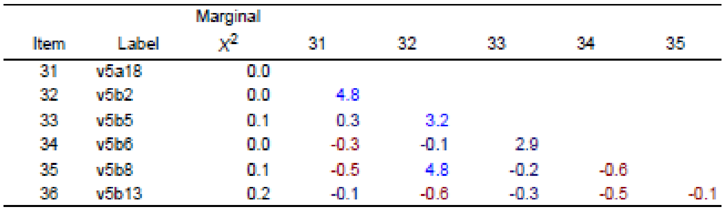
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

Item	Label	Marginal χ^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.0										
2	m3	0.0	8.1									
3	m1	0.0	2.9	-0.7								
4	m13	0.0	9.8	-0.5	2.4							
5	m18	0.0	-0.7	1.8	2.2	5.8						
6	m9	0.0	0.4	-0.6	2.0	0.5	-0.4					
7	m4	0.2	3.5	1.4	0.0	-0.5	1.9	-0.4				
8	m16	0.1	9.5	-0.5	0.9	-0.6	4.1	-0.5	0.7			
9	m14	0.2	7.0	0.6	5.6	-0.5	7.5	-0.5	1.0	0.7		
10	m6	0.0	4.2	2.9	-0.5	0.5	6.5	1.2	-0.4	1.4	-0.5	
11	m11	0.2	1.7	-0.4	1.1	1.8	2.1	-0.3	0.2	4.0	0.5	0.4
12	m7	0.3	3.6	1.5	0.3	0.4	0.1	-0.1	1.0	0.2	-0.3	0.2
13	m5	0.0	0.9	0.0	-0.7	5.5	-0.7	-0.6	-0.5	-0.5	7.2	-0.6
14	m2	0.0	1.1	3.5	-0.6	1.0	-0.7	0.4	7.8	2.0	5.4	4.8
15	m10	0.0	-0.7	2.0	-0.5	-0.3	-0.2	-0.7	5.3	4.4	7.7	8.8
16	m15	0.0	0.8	-0.3	0.6	0.1	1.0	-0.7	1.3	0.1	4.2	-0.2
17	m8	0.2	8.0	-0.5	-0.6	0.3	5.3	-0.3	0.2	0.0	1.2	-0.6
18	m17	0.1	1.5	3.5	0.9	0.1	2.7	-0.6	0.7	1.5	-0.5	-0.1
19	v3a1	0.0	9.3	2.2	-0.4	0.3	2.0	0.0	-0.4	-0.6	0.2	2.7
20	v3a3	0.1	10.2	1.6	-0.5	1.3	2.4	-0.1	-0.2	-0.5	0.1	-0.3
21	v3a9	0.1	-0.1	-0.3	-0.5	1.2	2.3	0.1	0.5	0.9	3.4	1.1
22	v3a17	0.1	9.9	-0.6	-0.5	-0.2	-0.1	2.6	2.8	-0.3	-0.4	-0.4
23	v3b4	0.1	4.9	0.4	-0.2	3.7	4.8	0.0	0.6	1.0	-0.0	-0.2
24	v3b12	0.2	26.3	0.6	-0.3	1.4	1.0	2.0	-0.1	0.9	-0.3	2.7
25	v3b15	0.1	-0.3	4.2	3.5	3.1	1.0	5.1	6.3	4.7	-0.0	3.5
26	v3b16	0.0	10.5	-0.3	0.7	0.6	1.6	0.8	2.4	-0.2	3.2	-0.5
27	v5a7	0.0	4.7	2.6	-0.5	-0.6	0.7	0.2	-0.6	-0.5	-0.5	-0.4
28	v5a10	0.0	2.2	-0.3	-0.5	-0.3	3.6	1.6	-0.6	-0.6	0.1	-0.3
29	v5a11	0.0	17.8	-0.7	1.3	-0.6	3.5	0.3	1.6	-0.2	-0.5	1.0
30	v5a14	0.2	0.3	1.7	-0.2	-0.5	1.2	0.4	0.4	-0.4	2.2	1.6
31	v5a18	0.0	0.3	0.2	1.7	-0.3	7.2	-0.1	-0.6	3.1	0.1	0.0
32	v5b2	0.0	4.3	0.2	-0.2	1.9	4.5	0.6	-0.7	-0.4	2.1	-0.0
33	v5b5	0.1	0.9	1.9	0.1	1.9	-0.5	2.6	-0.6	6.6	3.7	1.0
34	v5b6	0.0	7.8	-0.6	0.4	1.2	2.4	0.4	-0.4	0.7	-0.6	1.3
35	v5b8	0.1	0.9	-0.5	-0.1	-0.4	0.7	2.5	1.8	-0.4	-0.0	0.1
36	v5b13	0.2	2.2	-0.0	0.1	1.2	2.4	1.9	2.2	-0.3	-0.5	2.4

Item	Label	Marginal χ^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.2										
12	m7	0.3	-0.0									
13	m5	0.0	1.4	0.8								
14	m2	0.0	1.7	4.5	-0.4							
15	m10	0.0	1.6	0.1	2.3	1.1						
16	m15	0.0	9.4	0.5	-0.6	-0.7	0.5					
17	m8	0.2	2.2	-0.0	0.9	5.7	2.3	1.5				
18	m17	0.1	0.4	0.2	-0.7	3.2	4.4	0.1	-0.6			
19	v3a1	0.0	1.4	-0.6	2.1	-0.6	-0.5	-0.1	-0.7	-0.4		
20	v3a3	0.1	3.0	-0.5	-0.3	1.5	2.2	-0.4	0.5	1.5	0.6	
21	v3a9	0.1	0.5	-0.4	-0.0	1.0	0.7	-0.2	-0.3	-0.5	-0.6	0.6
22	v3a17	0.1	4.2	0.6	-0.2	7.2	2.0	0.8	2.3	-0.6	0.8	-0.4
23	v3b4	0.1	2.3	0.2	0.1	2.1	2.3	-0.0	0.9	-0.0	0.7	-0.6
24	v3b12	0.2	0.9	0.4	0.8	-0.3	1.0	1.2	0.4	2.7	-0.2	0.2
25	v3b15	0.1	0.1	11.1	13.9	0.2	1.7	7.2	2.6	3.5	4.1	4.3
26	v3b16	0.0	8.9	0.1	1.9	0.3	3.3	-0.2	-0.3	-0.1	-0.3	2.3
27	v5a7	0.0	4.5	0.2	0.8	1.6	-0.6	-0.1	-0.2	-0.4	1.1	-0.4
28	v5a10	0.0	0.9	0.3	-0.1	4.5	0.2	1.1	-0.2	1.1	1.0	1.7
29	v5a11	0.0	-0.6	0.9	3.5	0.1	4.7	5.9	3.2	0.6	0.6	-0.6
30	v5a14	0.2	1.6	-0.1	4.6	-0.1	0.3	-0.1	0.5	-0.3	0.7	5.2
31	v5a18	0.0	0.0	-0.2	3.3	-0.5	1.9	6.3	-0.3	0.3	2.9	2.9
32	v5b2	0.0	-0.6	-0.3	-0.2	0.0	0.6	1.1	1.2	-0.6	3.2	0.1
33	v5b5	0.1	-0.3	0.5	4.1	1.4	1.8	2.2	1.9	0.2	1.6	1.5
34	v5b6	0.0	4.3	-0.4	3.5	0.2	2.6	0.1	-0.5	-0.6	0.6	-0.6
35	v5b8	0.1	8.5	0.2	-0.4	1.1	0.0	-0.0	6.9	1.3	2.0	-0.4
36	v5b13	0.2	-0.0	2.3	-0.0	-0.4	0.0	-0.1	0.5	0.1	0.4	-0.5



Item	Label	Marginal χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.1										
22	v3a17	0.1	-0.6									
23	v3b4	0.1	0.6	-0.4								
24	v3b12	0.2	-0.2	-0.5	1.7							
25	v3b15	0.1	2.5	7.7	18.2	1.2						
26	v3b16	0.0	3.4	-0.5	1.7	0.7	9.0					
27	v5a7	0.0	-0.3	0.3	-0.4	-0.3	6.5	-0.5				
28	v5a10	0.0	-0.2	-0.5	1.3	0.7	8.0	1.5	-0.2			
29	v5a11	0.0	-0.3	0.2	-0.2	-0.5	3.0	-0.4	-0.3	0.4		
30	v5a14	0.2	0.8	-0.2	0.2	-0.5	3.0	-0.4	-0.3	-0.5	1.6	
31	v5a18	0.0	-0.2	-0.4	2.5	5.0	6.5	0.6	-0.2	1.4	1.3	-0.1
32	v5b2	0.0	2.3	1.2	0.2	-0.6	1.7	0.6	1.7	0.5	-0.6	-0.5
33	v5b5	0.1	0.2	-0.5	2.6	-0.4	4.6	1.3	1.2	0.1	-0.3	2.0
34	v5b6	0.0	-0.3	-0.4	-0.6	0.7	7.7	-0.6	3.8	-0.2	-0.7	1.3
35	v5b8	0.1	-0.6	6.9	-0.2	-0.5	6.4	0.3	-0.0	-0.1	1.1	-0.4
36	v5b13	0.2	-0.6	-0.2	-0.4	0.3	7.9	-0.5	-0.1	-0.5	-0.4	1.7



Item	Label	Marginal χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	4.8				
33	v5b5	0.1	0.3	3.2			
34	v5b6	0.0	-0.3	-0.1	2.9		
35	v5b8	0.1	-0.5	4.8	-0.2	-0.6	
36	v5b13	0.2	-0.1	-0.6	-0.3	-0.5	-0.1

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)

-2loglikelihood:	4521.93
Akaike Information Criterion (AIC):	4593.93
Bayesian Information Criterion (BIC):	4691.79

Summary of the Data and Control Parameters

Sample Size	112
Number of Items	36
Number of Dimensions	1

IRTPRO Version 4.2

Output generated by IRTPRO estimation engine Version 5.20 (64-bit)

Project: VT Sample - all items; Rasch

Item	*Label*	*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e.*
1	m12	1	-----	1	1.53	0.45	-1.53	0.45
2	m3	1	-----	2	-0.72	0.39	0.72	0.39
3	m1	1	-----	3	-0.72	0.39	0.72	0.39
4	m13	1	-----	4	-0.58	0.39	0.58	0.39
5	m18	1	-----	5	0.46	0.38	-0.46	0.38
6	m9	1	-----	6	0.2	0.38	-0.2	0.38
7	m4	1	-----	7	0.07	0.38	-0.07	0.38
8	m16	1	-----	8	0.52	0.39	-0.52	0.39
9	m14	1	-----	9	1.74	0.48	-1.74	0.48
10	m6	1	-----	10	-0.19	0.38	0.19	0.38
11	m11	1	-----	11	1.18	0.42	-1.18	0.42
12	m7	1	-----	12	0.46	0.38	-0.46	0.38
13	m5	1	-----	13	-1.26	0.44	1.26	0.44
14	m2	1	-----	14	-0.27	0.38	0.27	0.38
15	m10	1	-----	15	-0.87	0.4	0.87	0.4
16	m15	1	-----	16	0.4	0.39	-0.4	0.39
17	m8	1	-----	17	1.18	0.42	-1.18	0.42
18	m17	1	-----	18	0.2	0.38	-0.2	0.38
19	v3a1	1	-----	19	0.73	0.39	-0.73	0.39
20	v3a3	1	-----	20	-0.58	0.39	0.58	0.39
21	v3a9	1	-----	21	-0.45	0.38	0.45	0.38
22	v3a17	1	-----	22	0.73	0.39	-0.73	0.39
23	v3b4	1	-----	23	0.2	0.38	-0.2	0.38
24	v3b12	1	-----	24	3.02	0.74	-3.02	0.74
25	v3b15	1	-----	25	0.2	0.38	-0.2	0.38
26	v3b16	1	-----	26	-0.58	0.39	0.58	0.39
27	v5a7	1	-----	27	-0.45	0.38	0.45	0.38
28	v5a10	1	-----	28	-0.06	0.38	0.06	0.38
29	v5a11	1	-----	29	-0.32	0.38	0.32	0.38
30	v5a14	1	-----	30	1.35	0.44	-1.35	0.44
31	v5a18	1	-----	31	0.07	0.38	-0.07	0.38
32	v5b2	1	-----	32	0.33	0.38	-0.33	0.38
33	v5b5	1	-----	33	-0.32	0.38	0.32	0.38
34	v5b6	1	-----	34	0.2	0.38	-0.2	0.38
35	v5b8	1	-----	35	0.73	0.39	-0.73	0.39
36	v5b13	1	-----	36	1.02	0.41	-1.02	0.41

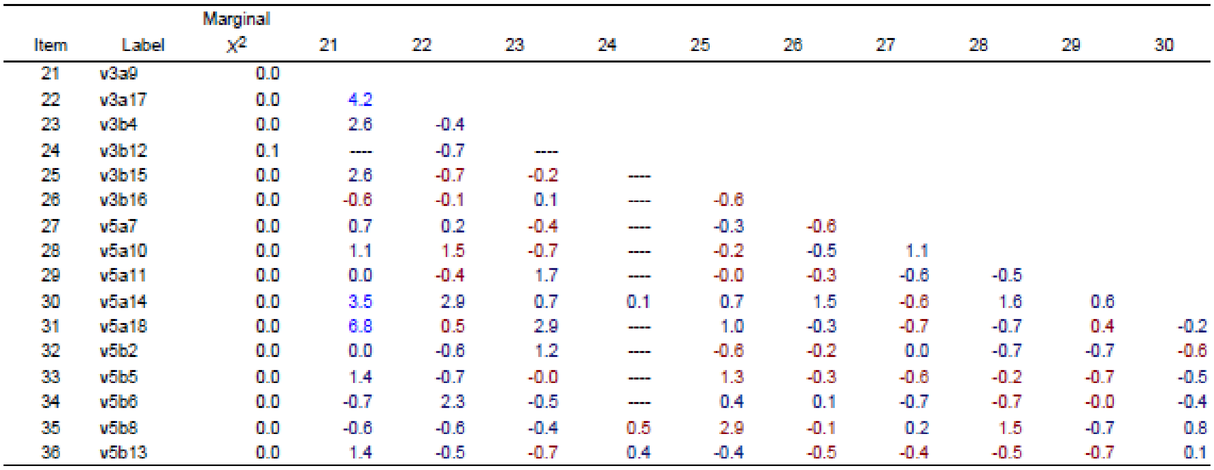
S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	*Probability*
1	m12	7.76	2	0.0207
2	m3	7.88	5	0.1625
3	m1	4.23	5	0.5173
4	m13	3.74	4	0.4427
5	m18	9.5	5	0.0905
6	m9	9.55	6	0.1447
7	m4	4.69	4	0.3217
8	m16	7.67	7	0.3642
9	m14	4.89	2	0.0866
10	m6	9.22	5	0.1003
11	m11	5.44	4	0.2468
12	m7	4.04	4	0.4021
13	m5	3.94	5	0.5595
14	m2	13.64	6	0.0338
15	m10	8.17	5	0.1466
16	m15	6.76	5	0.2405
17	m8	2.26	4	0.6895
18	m17	9.18	5	0.1017
19	v3a1	5.12	5	0.4026
20	v3a3	3.67	4	0.4541
21	v3a9	11.59	6	0.0716
22	v3a17	2.82	5	0.7289
23	v3b4	4.82	5	0.4393
24	v3b12	4.19	1	0.0405
25	v3b15	4.53	5	0.4766
26	v3b16	8.02	5	0.1547
27	v5a7	9.03	4	0.0602
28	v5a10	3.72	7	0.8119
29	v5a11	2.91	5	0.7149
30	v5a14	8.06	4	0.0891
31	v5a18	7.27	5	0.201
32	v5b2	8.3	5	0.1402
33	v5b5	6.14	4	0.1889
34	v5b6	12.86	6	0.0452
35	v5b8	4.46	5	0.4869
36	v5b13	9.91	5	0.0777

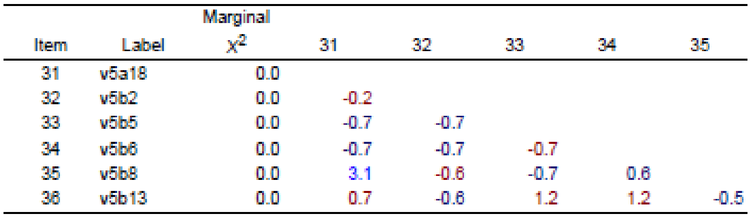
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

Item	Label	Marginal χ^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.1										
2	m3	0.0	0.7									
3	m1	0.0	-0.5	0.3								
4	m13	0.0	0.4	-0.5	-0.7							
5	m18	0.0	0.9	-0.6	0.2	-0.2						
6	m9	0.0	0.1	1.0	-0.2	0.1	2.1					
7	m4	0.0	0.3	1.5	-0.6	0.9	1.1	-0.7				
8	m16	0.0	2.9	0.0	-0.5	0.5	-0.1	0.2	-0.6			
9	m14	0.1	0.2	1.5	1.5	1.0	0.2	-0.4	-0.6	0.1		
10	m6	0.0	3.4	-0.4	-0.4	1.3	0.8	-0.7	-0.2	-0.7	0.1	
11	m11	0.0	0.8	1.3	1.3	3.0	0.6	0.0	-0.5	-0.5	0.4	-0.7
12	m7	0.0	-0.4	-0.6	0.2	-0.7	5.0	-0.4	0.3	-0.5	-0.6	-0.3
13	m5	0.1	-0.6	-0.5	-0.5	1.0	6.4	-0.6	-0.6	-0.3	----	-0.6
14	m2	0.0	1.3	-0.1	-0.5	-0.7	3.1	-0.6	-0.0	3.8	1.1	3.6
15	m10	0.0	-0.4	-0.7	-0.3	0.0	-0.7	-0.0	-0.3	-0.2	-0.1	0.0
16	m15	0.0	3.5	-0.7	-0.4	2.8	-0.1	0.2	-0.3	-0.2	-0.6	-0.4
17	m8	0.0	-0.6	-0.7	3.9	-0.6	-0.4	0.0	0.4	-0.6	-0.4	6.6
18	m17	0.0	-0.6	-0.7	-0.2	0.2	-0.4	-0.5	2.9	-0.7	0.8	-0.2
19	v3a1	0.0	-0.6	-0.3	-0.1	-0.7	-0.2	-0.4	-0.5	-0.4	-0.7	-0.4
20	v3a3	0.0	5.9	-0.1	-0.5	-0.7	-0.2	-0.6	-0.3	-0.6	13.5	-0.0
21	v3a9	0.0	0.1	-0.2	-0.7	-0.6	-0.7	2.6	0.3	0.8	6.9	-0.4
22	v3a17	0.0	0.2	-0.7	-0.7	-0.4	-0.2	0.6	-0.5	-0.5	-0.0	0.6
23	v3b4	0.0	4.4	1.0	-0.7	0.1	2.1	-0.5	1.8	-0.4	0.8	-0.7
24	v3b12	0.1	1.8	----	----	----	----	----	----	----	-0.4	----
25	v3b15	0.0	1.7	-0.4	-0.4	0.1	2.1	2.0	-0.5	0.6	-0.4	-0.7
26	v3b16	0.0	0.4	1.1	-0.5	-0.2	0.6	0.1	2.3	2.1	-0.4	-0.6
27	v5a7	0.0	0.1	-0.4	-0.7	0.0	0.5	-0.3	-0.5	-0.3	0.7	-0.7
28	v5a10	0.0	6.0	-0.7	-0.2	-0.5	-0.6	1.8	-0.2	-0.2	-0.1	1.0
29	v5a11	0.0	-0.1	-0.6	-0.6	-0.4	0.0	1.7	-0.2	-0.5	0.4	1.2
30	v5a14	0.0	-0.4	2.1	0.2	-0.7	0.1	2.8	1.1	-0.1	0.1	2.1
31	v5a18	0.0	-0.5	-0.6	-0.6	-0.4	-0.5	-0.2	0.3	-0.4	0.1	1.8
32	v5b2	0.0	-0.7	-0.4	-0.7	-0.7	-0.6	1.2	-0.7	-0.7	-0.5	0.2
33	v5b5	0.0	7.9	-0.6	-0.6	-0.4	0.0	3.8	-0.5	0.0	0.4	-0.6
34	v5b6	0.0	1.7	-0.7	0.6	-0.6	-0.4	-0.5	1.0	-0.2	3.2	-0.2
35	v5b8	0.0	0.2	-0.3	-0.1	0.7	-0.2	0.6	-0.2	4.1	-0.2	0.6
36	v5b13	0.0	1.2	0.2	0.7	-0.5	-0.5	-0.7	-0.7	-0.3	0.7	2.3

Item	Label	Marginal χ^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.0										
12	m7	0.0	-0.4									
13	m5	0.1	0.1	2.0								
14	m2	0.0	4.0	0.3	0.3							
15	m10	0.0	-0.6	0.7	1.9	-0.2						
16	m15	0.0	-0.1	2.9	0.0	0.7	-0.2					
17	m8	0.0	0.1	-0.0	-0.6	-0.4	0.1	-0.4				
18	m17	0.0	-0.5	-0.7	0.2	3.4	-0.5	0.8	0.0			
19	v3a1	0.0	-0.1	-0.7	-0.6	0.2	-0.6	-0.5	-0.4	0.6		
20	v3a3	0.0	0.1	-0.4	-0.6	-0.5	1.5	-0.7	-0.4	-0.6	0.7	
21	v3a9	0.0	5.2	4.6	1.9	-0.4	0.4	0.3	0.4	2.6	1.8	-0.6
22	v3a17	0.0	-0.7	-0.2	-0.6	0.2	-0.5	2.1	-0.4	0.6	0.3	-0.4
23	v3b4	0.0	-0.6	-0.3	-0.4	-0.4	-0.0	-0.5	1.6	0.4	-0.4	0.2
24	v3b12	0.1	0.3	----	----	----	----	----	-0.6	----	-0.7	----
25	v3b15	0.0	0.5	-0.3	-0.6	0.5	-0.5	-0.7	-0.5	-0.5	-0.4	0.1
26	v3b16	0.0	-0.4	-0.2	1.0	-0.5	-0.6	-0.3	0.1	0.1	-0.7	-0.7
27	v5a7	0.0	0.4	-0.2	-0.2	-0.2	4.5	1.9	-0.5	2.4	4.2	-0.6
28	v5a10	0.0	-0.7	1.8	3.1	0.6	-0.6	1.7	0.7	0.3	-0.6	-0.5
29	v5a11	0.0	0.7	-0.6	-0.4	3.9	0.9	0.1	-0.4	-0.7	-0.7	-0.3
30	v5a14	0.0	-0.2	-0.4	-0.1	0.5	0.5	8.1	-0.2	1.6	-0.4	1.5
31	v5a18	0.0	-0.5	-0.7	-0.1	6.0	0.8	-0.6	-0.5	0.3	-0.5	0.9
32	v5b2	0.0	3.2	-0.6	-0.5	-0.6	0.3	0.2	-0.3	-0.1	1.6	2.2
33	v5b5	0.0	0.7	-0.6	0.6	2.6	-0.3	-0.6	1.7	-0.0	-0.7	-0.7
34	v5b6	0.0	-0.6	0.5	0.2	1.3	3.5	0.0	1.6	-0.2	-0.7	-0.6
35	v5b8	0.0	-0.1	1.0	0.1	0.9	-0.6	-0.4	-0.4	0.6	0.1	0.7
36	v5b13	0.0	-0.6	-0.7	3.4	2.7	1.2	-0.6	-0.5	-0.1	-0.6	-0.5



Item	Label	Marginal χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.0										
22	v3a17	0.0	4.2									
23	v3b4	0.0	2.6	-0.4								
24	v3b12	0.1	----	-0.7	----							
25	v3b15	0.0	2.6	-0.7	-0.2	----						
26	v3b16	0.0	-0.6	-0.1	0.1	----	-0.6					
27	v5a7	0.0	0.7	0.2	-0.4	----	-0.3	-0.6				
28	v5a10	0.0	1.1	1.5	-0.7	----	-0.2	-0.5	1.1			
29	v5a11	0.0	0.0	-0.4	1.7	----	-0.0	-0.3	-0.6	-0.5		
30	v5a14	0.0	3.5	2.9	0.7	0.1	0.7	1.5	-0.6	1.6	0.6	
31	v5a18	0.0	6.8	0.5	2.9	----	1.0	-0.3	-0.7	-0.7	0.4	-0.2
32	v5b2	0.0	0.0	-0.6	1.2	----	-0.6	-0.2	0.0	-0.7	-0.7	-0.6
33	v5b5	0.0	1.4	-0.7	-0.0	----	1.3	-0.3	-0.6	-0.2	-0.7	-0.5
34	v5b6	0.0	-0.7	2.3	-0.5	----	0.4	0.1	-0.7	-0.7	-0.0	-0.4
35	v5b8	0.0	-0.6	-0.6	-0.4	0.5	2.9	-0.1	0.2	1.5	-0.7	0.8
36	v5b13	0.0	1.4	-0.5	-0.7	0.4	-0.4	-0.5	-0.4	-0.5	-0.7	0.1



Item	Label	Marginal χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	-0.2				
33	v5b5	0.0	-0.7	-0.7			
34	v5b6	0.0	-0.7	-0.7	-0.7		
35	v5b8	0.0	3.1	-0.6	-0.7	0.6	
36	v5b13	0.0	0.7	-0.6	1.2	1.2	-0.5

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	1554.66
Akaike Information Criterion (AIC):	1626.66
Bayesian Information Criterion (BIC):	1682.65

Summary of the Data and Control Parameters

Sample Size	35
Number of Items	36
Number of Dimensions	1

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: National Sample - all items; 1PL

Item	*Label*		*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e.*
1	m12	37	0.57	0.06	1	-0.27	0.2	0.47	0.35
2	m3	37	0.57	0.06	2	0.43	0.2	-0.75	0.36
3	m1	37	0.57	0.06	3	0.14	0.2	-0.25	0.35
4	m13	37	0.57	0.06	4	0.59	0.2	-1.04	0.37
5	m18	37	0.57	0.06	5	-0.06	0.2	0.1	0.34
6	m9	37	0.57	0.06	6	-0.35	0.2	0.61	0.35
7	m4	37	0.57	0.06	7	1.56	0.25	-2.74	0.51
8	m16	37	0.57	0.06	8	0.67	0.2	-1.18	0.38
9	m14	37	0.57	0.06	9	1.98	0.28	-3.48	0.6
10	m6	37	0.57	0.06	10	0.13	0.2	-0.24	0.35
11	m11	37	0.57	0.06	11	1.39	0.24	-2.44	0.48
12	m7	37	0.57	0.06	12	1.89	0.28	-3.31	0.58
13	m5	37	0.57	0.06	13	-0.04	0.2	0.07	0.34
14	m2	37	0.57	0.06	14	-0.1	0.2	0.17	0.34
15	m10	37	0.57	0.06	15	-0.08	0.2	0.13	0.34
16	m15	37	0.57	0.06	16	0.04	0.2	-0.07	0.34
17	m8	37	0.57	0.06	17	1.33	0.23	-2.34	0.47
18	m17	37	0.57	0.06	18	0.67	0.2	-1.18	0.38
19	v3a1	37	0.57	0.06	19	0.67	0.21	-1.18	0.39
20	v3a3	37	0.57	0.06	20	-0.23	0.21	0.41	0.36
21	v3a9	37	0.57	0.06	21	-0.42	0.21	0.73	0.38
22	v3a17	37	0.57	0.06	22	1.32	0.24	-2.33	0.48
23	v3b4	37	0.57	0.06	23	1.06	0.24	-1.86	0.45
24	v3b12	37	0.57	0.06	24	1.29	0.25	-2.27	0.49
25	v3b15	37	0.57	0.06	25	-0.82	0.23	1.44	0.43
26	v3b16	37	0.57	0.06	26	0.04	0.21	-0.08	0.37
27	v5a7	37	0.57	0.06	27	0.43	0.22	-0.75	0.4
28	v5a10	37	0.57	0.06	28	0.14	0.22	-0.25	0.38
29	v5a11	37	0.57	0.06	29	-0.05	0.22	0.08	0.38
30	v5a14	37	0.57	0.06	30	1.44	0.26	-2.52	0.53
31	v5a18	37	0.57	0.06	31	0.14	0.22	-0.25	0.38
32	v5b2	37	0.57	0.06	32	0.25	0.22	-0.44	0.39
33	v5b5	37	0.57	0.06	33	1.05	0.25	-1.84	0.47
34	v5b6	37	0.57	0.06	34	0.45	0.23	-0.79	0.41
35	v5b8	37	0.57	0.06	35	1.18	0.25	-2.07	0.49
36	v5b13	37	0.57	0.06	36	1.69	0.29	-2.96	0.59

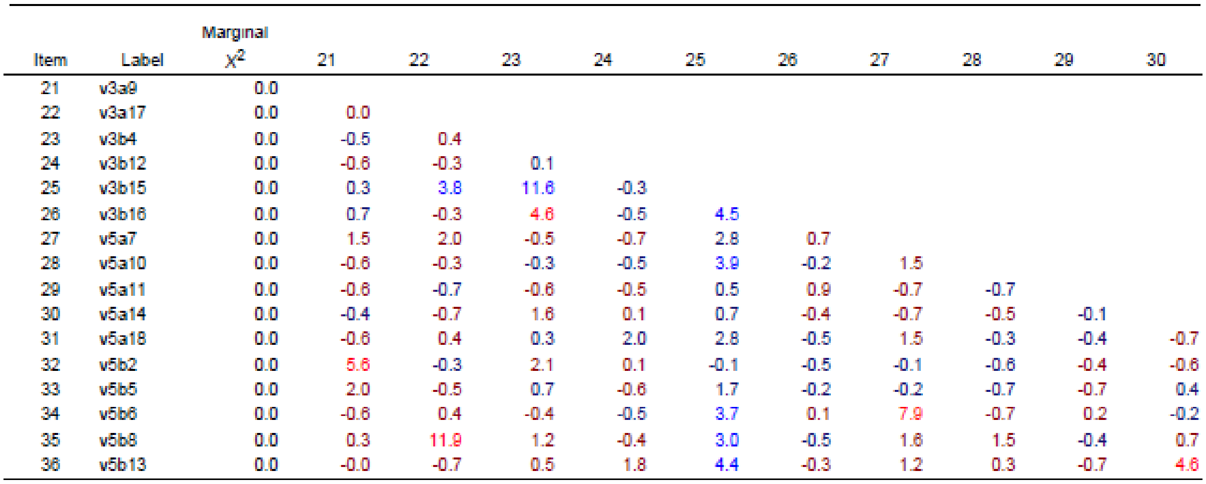
S-X^2 Item Level Diagnostic Statistics

Item	**Label**	***X^2***	***d.f.***	***Probability***
1	m12	31.13	13	0.0032
2	m3	15.6	11	0.1562
3	m1	10	12	0.6175
4	m13	11.4	11	0.4116
5	m18	19.18	14	0.1578
6	m9	12.54	12	0.4053
7	m4	18.96	7	0.0083
8	m16	21.84	11	0.0256
9	m14	11.7	7	0.1105
10	m6	11.23	12	0.5107
11	m11	13.84	8	0.0857
12	m7	9.43	6	0.1504
13	m5	8.65	10	0.567
14	m2	14.8	13	0.3218
15	m10	18.4	12	0.1038
16	m15	8.31	13	0.8232
17	m8	12.22	8	0.1415
18	m17	11.19	9	0.2622
19	v3a1	10.59	10	0.392
20	v3a3	14.15	12	0.2904
21	v3a9	10.76	10	0.3782
22	v3a17	16.52	8	0.0355
23	v3b4	10.1	8	0.2574
24	v3b12	19.64	8	0.0118
25	v3b15	47.56	11	0.0001
26	v3b16	5.94	13	0.9486
27	v5a7	16.03	10	0.0985
28	v5a10	17.58	10	0.0623
29	v5a11	17.84	12	0.1202
30	v5a14	7.64	7	0.3665
31	v5a18	8.89	13	0.7818
32	v5b2	7.66	12	0.8115
33	v5b5	11.11	8	0.1952
34	v5b6	14.19	11	0.2223
35	v5b8	11.12	9	0.2669
36	v5b13	10.97	6	0.089

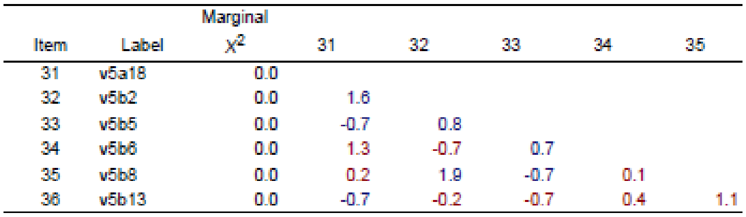
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

Item	Label	Marginal X^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.0										
2	m3	0.0	3.4									
3	m1	0.0	0.2	0.3								
4	m13	0.0	4.6	0.9	0.0							
5	m18	0.0	0.1	-0.3	-0.1	1.9						
6	m9	0.0	-0.7	-0.3	5.5	-0.7	-0.5					
7	m4	0.0	0.9	4.1	-0.7	0.1	-0.0	0.5				
8	m16	0.0	4.5	0.8	-0.6	-0.2	0.9	0.8	2.9			
9	m14	0.0	3.7	-0.5	2.8	-0.1	4.1	-0.3	3.3	2.7		
10	m6	0.0	0.9	0.2	-0.5	-0.7	2.3	-0.5	0.7	4.7	-0.5	
11	m11	0.0	-0.2	0.7	3.7	-0.0	0.0	-0.6	-0.7	1.4	-0.5	-0.7
12	m7	0.0	1.1	3.9	2.0	-0.6	-0.7	1.1	2.9	1.9	-0.7	-0.7
13	m5	0.0	-0.6	-0.7	0.5	1.7	-0.2	-0.3	-0.5	-0.5	3.9	0.8
14	m2	0.0	-0.5	0.5	-0.2	-0.5	-0.0	-0.7	3.9	-0.2	2.5	1.3
15	m10	0.0	0.5	-0.2	-0.5	-0.6	-0.6	-0.1	2.1	1.1	4.2	3.8
16	m15	0.0	-0.6	-0.6	-0.6	-0.7	-0.6	0.3	-0.3	-0.7	1.7	-0.6
17	m8	0.0	3.8	-0.3	-0.2	-0.7	1.9	0.8	1.8	1.8	3.7	-0.2
18	m17	0.0	-0.4	0.6	-0.6	-0.7	0.2	0.1	2.9	4.8	0.1	-0.7
19	v3a1	0.0	4.5	0.1	1.3	-0.6	-0.1	-0.1	-0.5	-0.4	-0.6	0.3
20	v3a3	0.0	5.0	4.8	-0.3	-0.4	0.1	0.2	-0.5	1.1	-0.7	-0.6
21	v3a9	0.0	-0.6	-0.6	0.0	-0.4	0.0	-0.1	2.5	-0.5	1.1	-0.4
22	v3a17	0.0	5.4	-0.4	-0.2	-0.7	-0.7	5.0	6.9	1.3	-0.1	-0.6
23	v3b4	0.0	1.7	-0.6	-0.7	1.1	1.7	0.8	2.9	3.8	0.7	-0.7
24	v3b12	0.0	18.7	-0.6	-0.7	-0.1	-0.4	0.1	-0.7	-0.4	-0.5	0.5
25	v3b15	0.0	-0.7	1.2	0.8	0.7	-0.5	2.0	3.1	1.6	-0.3	0.8
26	v3b16	0.0	5.4	-0.6	-0.6	-0.5	-0.2	-0.3	0.5	1.8	1.1	-0.6
27	v5a7	0.0	1.5	6.2	0.1	0.0	-0.5	1.4	0.5	-0.1	-0.2	-0.6
28	v5a10	0.0	0.1	-0.6	0.7	-0.5	0.8	4.1	-0.3	0.2	-0.7	-0.6
29	v5a11	0.0	11.0	-0.1	-0.3	0.7	0.8	-0.3	0.1	-0.4	-0.4	-0.5
30	v5a14	0.0	-0.6	0.0	-0.5	-0.1	-0.2	-0.3	2.5	0.2	5.0	-0.0
31	v5a18	0.0	-0.7	-0.7	-0.1	-0.5	12.3	0.1	-0.3	0.7	-0.7	-0.7
32	v5b2	0.0	1.3	-0.7	1.1	0.1	1.5	0.2	0.3	0.0	0.5	-0.7
33	v5b5	0.0	-0.3	0.1	-0.2	0.2	-0.1	0.7	-0.2	3.4	2.1	-0.4
34	v5b6	0.0	3.6	-0.5	-0.3	4.1	0.3	0.9	1.1	3.4	-0.1	4.1
35	v5b8	0.0	-0.4	-0.6	-0.3	-0.2	-0.4	4.5	5.1	0.0	1.2	-0.7
36	v5b13	0.0	0.4	1.4	1.1	3.6	0.6	3.3	5.6	0.1	-0.6	0.5

Item	Label	Marginal X^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.0										
12	m7	0.0	-0.7									
13	m5	0.0	-0.3	-0.5								
14	m2	0.0	-0.2	1.8	-0.5							
15	m10	0.0	-0.2	-0.7	-0.1	-0.5						
16	m15	0.0	5.0	-0.6	-0.3	0.1	-0.7					
17	m8	0.0	0.4	1.2	-0.5	2.2	0.1	-0.3				
18	m17	0.0	-0.6	1.9	-0.1	0.4	1.1	-0.7	-0.1			
19	v3a1	0.0	-0.0	0.1	0.1	-0.3	0.9	-0.6	0.3	1.2		
20	v3a3	0.0	0.7	-0.6	1.0	-0.3	-0.1	-0.4	-0.6	-0.3	-0.6	
21	v3a9	0.0	-0.4	0.5	-0.0	-0.5	-0.6	-0.1	-0.6	-0.4	0.1	-0.6
22	v3a17	0.0	2.1	-0.4	0.5	3.5	0.0	-0.4	6.0	-0.1	-0.5	0.8
23	v3b4	0.0	0.6	0.7	0.1	0.1	0.2	1.2	3.1	1.6	2.6	0.1
24	v3b12	0.0	-0.1	-0.5	-0.2	-0.7	-0.4	-0.0	-0.6	0.6	-0.4	-0.7
25	v3b15	0.0	2.0	6.9	8.5	-0.6	-0.1	3.3	0.4	0.8	1.2	1.4
26	v3b16	0.0	4.9	-0.5	0.1	-0.7	0.6	0.4	-0.6	-0.7	0.2	5.8
27	v5a7	0.0	2.0	1.0	-0.3	-0.3	-0.1	1.0	0.4	-0.6	-0.1	1.0
28	v5a10	0.0	-0.0	1.2	-0.3	1.3	2.3	-0.1	-0.4	-0.4	-0.1	-0.2
29	v5a11	0.0	0.5	-0.4	0.8	-0.7	1.5	2.4	0.8	-0.6	-0.1	-0.2
30	v5a14	0.0	0.5	-0.2	1.8	-0.7	-0.6	0.2	-0.5	-0.7	2.0	2.1
31	v5a18	0.0	-0.2	-0.5	0.7	-0.6	-0.1	2.7	0.1	-0.7	0.7	0.4
32	v5b2	0.0	0.8	0.2	-0.5	2.0	-0.4	-0.0	-0.3	0.1	6.2	-0.7
33	v5b5	0.0	-0.4	-0.5	7.7	-0.2	0.2	0.4	0.3	-0.7	0.1	-0.2
34	v5b6	0.0	1.9	-0.1	0.8	-0.7	0.4	0.0	-0.3	-0.3	-0.2	0.6
35	v5b8	0.0	5.8	-0.6	0.2	-0.3	-0.5	0.1	12.0	-0.2	0.4	-0.6
36	v5b13	0.0	-0.5	4.7	-0.5	-0.6	1.3	0.3	-0.5	1.5	1.4	-0.6



Item	Label	Marginal χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.0										
22	v3a17	0.0	0.0									
23	v3b4	0.0	-0.5	0.4								
24	v3b12	0.0	-0.6	-0.3	0.1							
25	v3b15	0.0	0.3	3.8	11.6	-0.3						
26	v3b16	0.0	0.7	-0.3	4.6	-0.5	4.5					
27	v5a7	0.0	1.5	2.0	-0.5	-0.7	2.8	0.7				
28	v5a10	0.0	-0.6	-0.3	-0.3	-0.5	3.9	-0.2	1.5			
29	v5a11	0.0	-0.6	-0.7	-0.6	-0.5	0.5	0.9	-0.7	-0.7		
30	v5a14	0.0	-0.4	-0.7	1.6	0.1	0.7	-0.4	-0.7	-0.5	-0.1	
31	v5a18	0.0	-0.6	0.4	0.3	2.0	2.8	-0.5	1.5	-0.3	-0.4	-0.7
32	v5b2	0.0	5.6	-0.3	2.1	0.1	-0.1	-0.5	-0.1	-0.6	-0.4	-0.6
33	v5b5	0.0	2.0	-0.5	0.7	-0.6	1.7	-0.2	-0.2	-0.7	-0.7	0.4
34	v5b6	0.0	-0.6	0.4	-0.4	-0.5	3.7	0.1	7.9	-0.7	0.2	-0.2
35	v5b8	0.0	0.3	11.9	1.2	-0.4	3.0	-0.5	1.6	1.5	-0.4	0.7
36	v5b13	0.0	-0.0	-0.7	0.5	1.8	4.4	-0.3	1.2	0.3	-0.7	4.6



Item	Label	Marginal χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	1.6				
33	v5b5	0.0	-0.7	0.8			
34	v5b6	0.0	1.3	-0.7	0.7		
35	v5b8	0.0	0.2	1.9	-0.7	0.1	
36	v5b13	0.0	-0.7	-0.2	-0.7	0.4	1.1

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	4491.3
Akaike Information Criterion (AIC):	4565.3
Bayesian Information Criterion (BIC):	4665.88

Summary of the Data and Control Parameters

Sample Size	112
Number of Items	36
Number of Dimensions	1

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: VT Sample - all items; 1PL

Item	*Label*		*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e.*
1	m12	36	0.6	0.11	1	1.49	0.43	-2.46	0.82
2	m3	36	0.6	0.11	2	-0.7	0.36	1.16	0.63
3	m1	36	0.6	0.11	3	-0.7	0.36	1.16	0.63
4	m13	36	0.6	0.11	4	-0.57	0.36	0.94	0.61
5	m18	36	0.6	0.11	5	0.44	0.35	-0.73	0.59
6	m9	36	0.6	0.11	6	0.19	0.35	-0.31	0.58
7	m4	36	0.6	0.11	7	0.06	0.35	-0.11	0.57
8	m16	36	0.6	0.11	8	0.5	0.36	-0.83	0.61
9	m14	36	0.6	0.11	9	1.69	0.46	-2.79	0.89
10	m6	36	0.6	0.11	10	-0.18	0.35	0.3	0.58
11	m11	36	0.6	0.11	11	1.14	0.4	-1.89	0.73
12	m7	36	0.6	0.11	12	0.44	0.35	-0.73	0.59
13	m5	36	0.6	0.11	13	-1.23	0.42	2.04	0.77
14	m2	36	0.6	0.11	14	-0.26	0.35	0.43	0.59
15	m10	36	0.6	0.11	15	-0.84	0.37	1.39	0.66
16	m15	36	0.6	0.11	16	0.39	0.36	-0.64	0.6
17	m8	36	0.6	0.11	17	1.14	0.4	-1.89	0.73
18	m17	36	0.6	0.11	18	0.19	0.35	-0.31	0.58
19	v3a1	36	0.6	0.11	19	0.71	0.36	-1.17	0.63
20	v3a3	36	0.6	0.11	20	-0.57	0.36	0.94	0.61
21	v3a9	36	0.6	0.11	21	-0.44	0.35	0.72	0.6
22	v3a17	36	0.6	0.11	22	0.71	0.36	-1.17	0.63
23	v3b4	36	0.6	0.11	23	0.19	0.35	-0.31	0.58
24	v3b12	36	0.6	0.11	24	2.96	0.69	-4.89	1.39
25	v3b15	36	0.6	0.11	25	0.19	0.35	-0.31	0.58
26	v3b16	36	0.6	0.11	26	-0.57	0.36	0.94	0.61
27	v5a7	36	0.6	0.11	27	-0.44	0.35	0.72	0.6
28	v5a10	36	0.6	0.11	28	-0.06	0.35	0.1	0.57
29	v5a11	36	0.6	0.11	29	-0.31	0.35	0.51	0.58
30	v5a14	36	0.6	0.11	30	1.31	0.42	-2.17	0.77
31	v5a18	36	0.6	0.11	31	0.06	0.35	-0.11	0.57
32	v5b2	36	0.6	0.11	32	0.31	0.35	-0.52	0.58
33	v5b5	36	0.6	0.11	33	-0.31	0.35	0.51	0.58
34	v5b6	36	0.6	0.11	34	0.19	0.35	-0.31	0.58
35	v5b8	36	0.6	0.11	35	0.71	0.36	-1.17	0.63
36	v5b13	36	0.6	0.11	37	0.99	0.39	-1.64	0.7

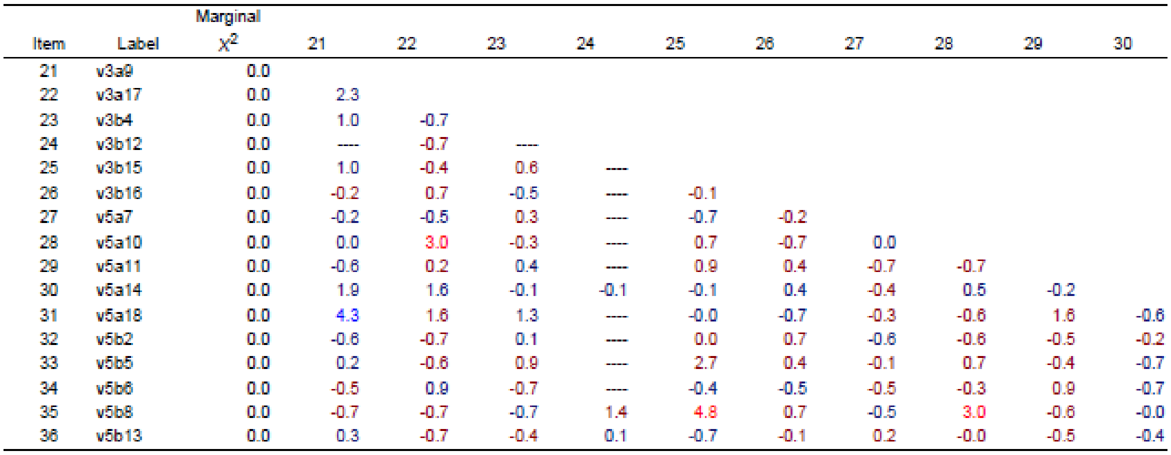
S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	*Probability*
1	m12	6.49	1	0.0108
2	m3	8.67	4	0.0698
3	m1	4.33	4	0.3639
4	m13	3.05	3	0.3855
5	m18	8.02	3	0.0456
6	m9	8.89	5	0.1132
7	m4	4.35	4	0.3618
8	m16	7.74	4	0.1013
9	m14	4.56	2	0.1021
10	m6	9.87	3	0.0196
11	m11	5.4	3	0.1446
12	m7	4.07	3	0.2552
13	m5	3.87	4	0.4254
14	m2	11.31	5	0.0455
15	m10	6.97	3	0.0727
16	m15	7.55	5	0.1822
17	m8	4.18	4	0.3834
18	m17	10.18	5	0.0701
19	v3a1	5.53	4	0.2383
20	v3a3	6.23	3	0.1009
21	v3a9	9.09	4	0.0587
22	v3a17	3.03	4	0.5543
23	v3b4	3.81	5	0.5785
24	v3b12			
25	v3b15	4.96	5	0.4218
26	v3b16	8.89	4	0.0639
27	v5a7	9.58	4	0.0481
28	v5a10	3.68	6	0.7203
29	v5a11	3.25	4	0.5183
30	v5a14	6.19	2	0.0452
31	v5a18	9.26	5	0.0989
32	v5b2	9.2	4	0.0562
33	v5b5	7.17	3	0.0665
34	v5b6	9.18	5	0.1019
35	v5b8	4.47	4	0.3476
36	v5b13	7.3	3	0.0628

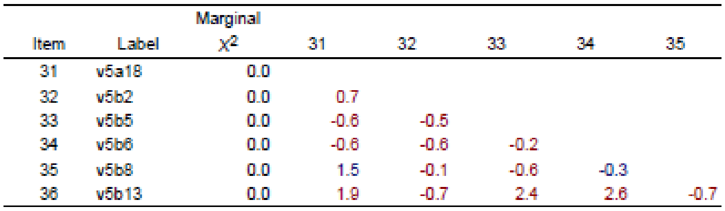
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

Item	Label	Marginal X^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.0										
2	m3	0.0	-0.1									
3	m1	0.0	-0.7	1.4								
4	m13	0.0	-0.3	-0.7	-0.4							
5	m18	0.0	0.1	-0.7	-0.5	-0.6						
6	m9	0.0	-0.5	2.3	0.6	-0.5	0.7					
7	m4	0.0	-0.3	3.0	-0.2	-0.1	0.0	-0.6				
8	m16	0.0	1.7	0.9	-0.7	-0.3	0.8	-0.5	-0.7			
9	m14	0.0	1.1	0.4	0.4	0.1	-0.3	-0.7	-0.4	-0.4		
10	m6	0.0	1.9	-0.7	-0.7	0.2	-0.2	-0.6	0.7	-0.6	-0.5	
11	m11	0.0	0.1	0.2	0.2	1.5	-0.2	-0.5	-0.7	-0.5	-0.2	-0.6
12	m7	0.0	-0.7	-0.1	-0.5	-0.6	2.9	-0.7	1.5	-0.7	-0.7	-0.7
13	m5	0.0	-0.6	-0.2	-0.2	0.1	4.3	-0.6	-0.6	-0.7	-0.7	-0.4
14	m2	0.0	0.3	-0.6	-0.7	-0.5	1.5	-0.6	-0.6	2.0	0.2	1.8
15	m10	0.0	-0.7	-0.5	0.4	-0.5	-0.4	-0.6	-0.7	-0.6	-0.6	-0.6
16	m15	0.0	2.1	-0.6	-0.7	1.2	0.8	-0.5	-0.7	0.5	-0.7	-0.7
17	m8	0.0	-0.3	-0.5	2.1	-0.4	-0.7	-0.5	-0.4	-0.2	0.2	4.4
18	m17	0.0	-0.6	-0.4	0.6	1.2	-0.7	-0.7	1.3	-0.5	0.0	-0.7
19	v3a1	0.0	-0.7	0.3	-0.6	-0.4	-0.6	-0.7	0.2	-0.7	-0.5	-0.7
20	v3a3	0.0	3.8	0.7	-0.7	-0.6	-0.6	-0.7	-0.7	-0.6	10.4	-0.6
21	v3a9	0.0	-0.5	-0.6	-0.6	-0.7	-0.4	1.0	-0.5	-0.1	4.8	-0.7
22	v3a17	0.0	-0.4	-0.6	-0.6	-0.7	-0.6	-0.3	0.2	-0.6	0.8	-0.2
23	v3b4	0.0	2.8	2.3	-0.4	-0.5	0.7	-0.7	3.4	0.3	0.0	-0.6
24	v3b12	0.0	3.7	----	----	----	-0.7	----	----	-0.7	-0.0	----
25	v3b15	0.0	0.7	-0.7	-0.7	-0.5	0.7	0.6	0.1	1.8	-0.7	-0.6
26	v3b16	0.0	-0.3	2.6	-0.7	-0.6	1.8	-0.5	4.0	3.7	-0.7	-0.6
27	v5a7	0.0	-0.5	0.2	-0.6	1.0	-0.3	-0.7	-0.7	0.4	-0.1	-0.4
28	v5a10	0.0	4.1	-0.6	-0.6	-0.7	-0.7	0.5	-0.6	0.6	-0.6	-0.0
29	v5a11	0.0	0.6	-0.7	-0.7	-0.7	-0.6	0.4	-0.7	-0.7	-0.3	0.1
30	v5a14	0.0	0.0	0.8	-0.4	-0.6	-0.5	1.4	0.2	-0.6	-0.4	0.9
31	v5a18	0.0	-0.0	-0.7	-0.2	0.3	0.1	-0.6	-0.4	-0.7	0.9	0.5
32	v5b2	0.0	-0.5	0.2	-0.6	-0.4	-0.7	0.1	-0.3	-0.5	-0.7	-0.5
33	v5b5	0.0	5.6	-0.1	-0.7	-0.7	-0.6	1.9	0.2	0.9	-0.3	0.0
34	v5b6	0.0	0.7	-0.4	-0.2	-0.7	-0.7	-0.7	-0.0	-0.6	1.9	-0.7
35	v5b8	0.0	-0.4	0.3	-0.6	-0.2	-0.6	-0.3	-0.6	6.2	-0.6	-0.2
36	v5b13	0.0	0.4	1.0	-0.2	-0.1	0.1	-0.4	-0.6	0.4	0.0	0.9

Item	Label	Marginal X^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.0										
12	m7	0.0	-0.7									
13	m5	0.0	-0.5	3.3								
14	m2	0.0	2.2	-0.4	-0.3							
15	m10	0.0	-0.6	-0.2	0.9	-0.6						
16	m15	0.0	0.7	1.3	-0.5	-0.2	0.4					
17	m8	0.0	-0.4	0.8	-0.3	0.3	-0.5	-0.7				
18	m17	0.0	0.1	-0.4	-0.4	1.6	0.1	-0.2	-0.5			
19	v3a1	0.0	-0.6	-0.6	-0.6	-0.5	-0.7	0.1	0.2	-0.3		
20	v3a3	0.0	0.9	0.2	-0.6	-0.7	0.4	-0.5	-0.7	-0.7	-0.2	
21	v3a9	0.0	3.2	2.6	0.8	0.3	-0.3	-0.4	-0.3	1.0	0.5	-0.2
22	v3a17	0.0	-0.6	-0.6	-0.6	-0.5	-0.0	3.8	0.2	-0.3	1.4	-0.7
23	v3b4	0.0	-0.7	0.6	0.1	-0.7	-0.6	-0.7	0.4	-0.4	-0.7	1.2
24	v3b12	0.0	0.0	0.7	----	----	----	----	-0.5	----	-0.7	----
25	v3b15	0.0	1.7	0.6	-0.6	-0.3	0.1	-0.5	0.1	-0.7	-0.7	-0.5
26	v3b16	0.0	-0.7	-0.6	0.1	-0.7	-0.7	0.5	0.9	-0.5	-0.4	-0.6
27	v5a7	0.0	-0.3	0.7	0.3	-0.6	2.7	0.6	-0.7	4.1	2.3	-0.2
28	v5a10	0.0	-0.4	0.5	1.6	-0.2	-0.3	3.2	-0.1	-0.4	-0.7	-0.7
29	v5a11	0.0	1.7	-0.7	0.0	2.1	-0.1	1.0	0.1	-0.2	-0.6	0.4
30	v5a14	0.0	-0.6	0.1	-0.5	-0.3	-0.2	6.1	-0.6	3.0	-0.7	0.4
31	v5a18	0.0	-0.7	-0.6	-0.6	3.7	-0.1	-0.0	-0.7	1.4	0.2	-0.1
32	v5b2	0.0	1.7	-0.1	-0.7	-0.6	-0.4	-0.5	0.4	-0.6	3.1	0.8
33	v5b5	0.0	1.7	-0.7	-0.2	1.1	-0.7	-0.2	0.5	0.9	-0.6	-0.5
34	v5b6	0.0	-0.7	-0.3	-0.4	0.2	1.8	1.0	0.4	0.6	-0.4	-0.1
35	v5b8	0.0	-0.6	0.0	-0.5	-0.0	-0.7	-0.7	0.2	-0.3	-0.5	-0.2
36	v5b13	0.0	-0.2	-0.6	1.8	1.2	0.2	-0.4	-0.7	0.7	-0.2	-0.1



Item	Label	Marginal χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.0										
22	v3a17	0.0	2.3									
23	v3b4	0.0	1.0	-0.7								
24	v3b12	0.0	----	-0.7	----							
25	v3b15	0.0	1.0	-0.4	0.6	----						
26	v3b16	0.0	-0.2	0.7	-0.5	----	-0.1					
27	v5a7	0.0	-0.2	-0.5	0.3	----	-0.7	-0.2				
28	v5a10	0.0	0.0	3.0	-0.3	----	0.7	-0.7	0.0			
29	v5a11	0.0	-0.6	0.2	0.4	----	0.9	0.4	-0.7	-0.7		
30	v5a14	0.0	1.9	1.6	-0.1	-0.1	-0.1	0.4	-0.4	0.5	-0.2	
31	v5a18	0.0	4.3	1.6	1.3	----	-0.0	-0.7	-0.3	-0.6	1.6	-0.6
32	v5b2	0.0	-0.6	-0.7	0.1	----	0.0	0.7	-0.6	-0.6	-0.5	-0.2
33	v5b5	0.0	0.2	-0.6	0.9	----	2.7	0.4	-0.1	0.7	-0.4	-0.7
34	v5b6	0.0	-0.5	0.9	-0.7	----	-0.4	-0.5	-0.5	-0.3	0.9	-0.7
35	v5b8	0.0	-0.7	-0.7	-0.7	1.4	4.8	0.7	-0.5	3.0	-0.6	-0.0
36	v5b13	0.0	0.3	-0.7	-0.4	0.1	-0.7	-0.1	0.2	-0.0	-0.5	-0.4



Item	Label	Marginal χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	0.7				
33	v5b5	0.0	-0.6	-0.5			
34	v5b6	0.0	-0.6	-0.6	-0.2		
35	v5b8	0.0	1.5	-0.1	-0.6	-0.3	
36	v5b13	0.0	1.9	-0.7	2.4	2.6	-0.7

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	1546.21
Akaike Information Criterion (AIC):	1620.21
Bayesian Information Criterion (BIC):	1677.76

Summary of the Data and Control Parameters

Sample Size	35
Number of Items	36
Number of Dimensions	1

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: National Sample - all items; 2PL

Item	***Label***		***a***	***s.e.***		***c***	***s.e.***	***b***	***s.e.***
1	m12	2	-	0.25	1	-0.28	0.21	-0.44	0.36
2	m3	4	0.93	0.3	3	0.48	0.24	-0.52	0.27
3	m1	6	0.74	0.27	5	0.16	0.21	-0.21	0.29
4	m13	8	0.52	0.25	7	0.59	0.21	-1.13	0.61
5	m18	10	-	0.21	9	-0.05	0.19	-0.48	1.89
6	m9	12	1.24	0.35	11	-0.4	0.26	0.32	0.22
7	m4	14	2.26	0.72	13	2.53	0.67	-1.12	0.2
8	m16	16	1.38	0.38	15	0.87	0.3	-0.63	0.21
9	m14	18	0.71	0.38	17	2.04	0.34	-2.87	1.33
10	m6	20	0.54	0.24	19	0.14	0.2	-0.25	0.38
11	m11	22	-	0.26	21	1.3	0.23	13.87	37.9
12	m7	24	1.62	0.58	23	2.47	0.59	-1.53	0.34
13	m5	26	0.48	0.23	25	-0.04	0.2	0.07	0.41
14	m2	28	-	0.21	27	-0.09	0.19	-3.19	24.82
15	m10	30	0.07	0.21	29	-0.07	0.19	1.05	4.31
16	m15	32	0.48	0.23	31	0.04	0.2	-0.08	0.41
17	m8	34	1.31	0.43	33	1.64	0.39	-1.25	0.3
18	m17	36	0.87	0.3	35	0.74	0.24	-0.84	0.33
19	v3a1	38	0.72	0.28	37	0.72	0.23	-0.99	0.41
20	v3a3	40	0.79	0.28	39	-0.22	0.23	0.28	0.31
21	v3a9	42	0.6	0.26	41	-0.41	0.21	0.68	0.46
22	v3a17	44	1.71	0.56	43	1.9	0.52	-1.11	0.23
23	v3b4	46	1.44	0.46	45	1.39	0.4	-0.96	0.24
24	v3b12	48	0.38	0.29	47	1.25	0.25	-3.3	2.46
25	v3b15	50	-	0.33	49	-0.9	0.28	-0.94	0.34
26	v3b16	52	0.73	0.28	51	0.06	0.22	-0.08	0.3
27	v5a7	54	1.38	0.4	53	0.59	0.32	-0.43	0.21
28	v5a10	56	0.7	0.28	55	0.16	0.22	-0.23	0.32
29	v5a11	58	0.4	0.24	57	-0.04	0.21	0.1	0.53
30	v5a14	60	0.69	0.34	59	1.49	0.3	-2.14	0.95
31	v5a18	62	0.3	0.24	61	0.14	0.21	-0.45	0.77
32	v5b2	64	0.51	0.26	63	0.26	0.22	-0.51	0.47
33	v5b5	66	0.11	0.27	65	0.97	0.24	-8.83	21.3
34	v5b6	68	1.17	0.37	67	0.59	0.3	-0.5	0.24
35	v5b8	70	1.58	0.57	69	1.64	0.52	-1.04	0.24
36	v5b13	72	1.96	0.85	71	2.57	0.86	-1.31	0.26

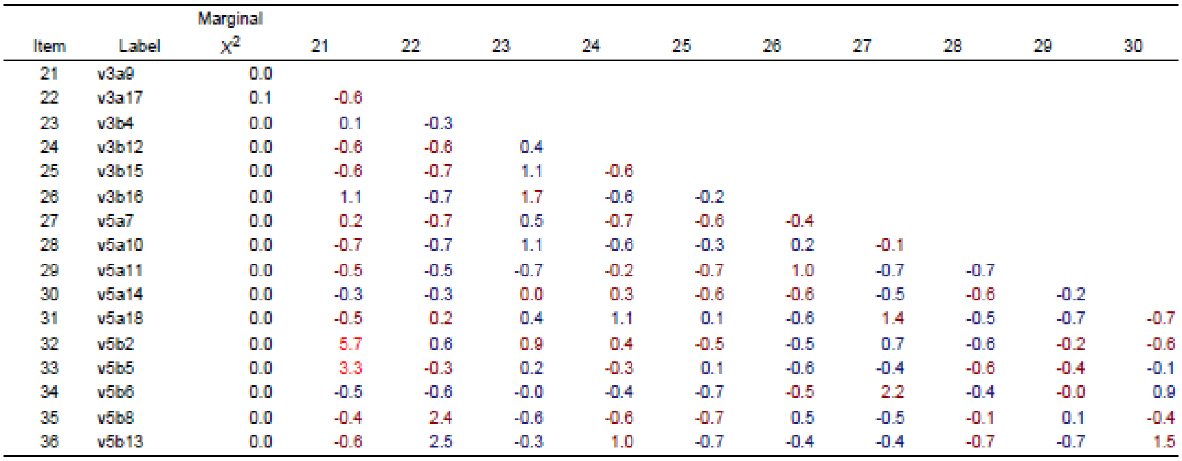
S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	*Probability*
1	m12	7.91	11	0.7219
2	m3	15.73	12	0.2033
3	m1	9.12	12	0.6938
4	m13	11.37	10	0.331
5	m18	16.54	14	0.2805
6	m9	7.71	12	0.8079
7	m4	9.24	5	0.0998
8	m16	19.06	10	0.0394
9	m14	8.53	5	0.1291
10	m6	11.04	12	0.5265
11	m11	10.2	8	0.2504
12	m7	8.37	5	0.1368
13	m5	8.57	10	0.5749
14	m2	15.76	13	0.2615
15	m10	16.78	12	0.1575
16	m15	7.89	13	0.8511
17	m8	10.57	7	0.1581
18	m17	9.99	9	0.353
19	v3a1	9.64	9	0.382
20	v3a3	15.73	13	0.2632
21	v3a9	10.87	10	0.3696
22	v3a17	14.07	7	0.0498
23	v3b4	8.47	7	0.2949
24	v3b12	17.47	8	0.0255
25	v3b15	17.8	11	0.0861
26	v3b16	6.7	13	0.9174
27	v5a7	11.13	9	0.2665
28	v5a10	17.33	10	0.0671
29	v5a11	16.96	13	0.2008
30	v5a14	7.93	7	0.3406
31	v5a18	11.42	14	0.6541
32	v5b2	8.71	10	0.5608
33	v5b5	8.11	12	0.7772
34	v5b6	9.67	9	0.3796
35	v5b8	9.11	8	0.3347
36	v5b13	3.68	5	0.5971

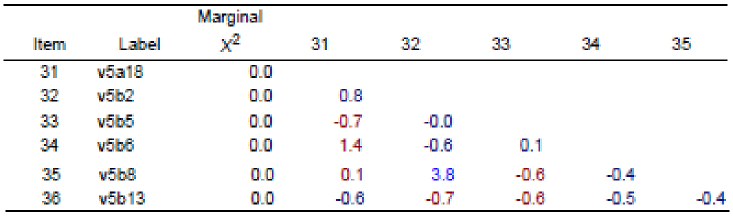
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

Item	Label	Marginal X^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.0										
2	m3	0.0	-0.5									
3	m1	0.0	-0.4	-0.5								
4	m13	0.0	0.4	0.4	0.2							
5	m18	0.0	1.1	-0.7	-0.7	0.1						
6	m9	0.0	2.2	-0.6	2.3	-0.3	0.9					
7	m4	0.0	-0.2	-0.1	0.7	-0.7	-0.7	-0.2				
8	m16	0.0	-0.6	-0.7	0.8	-0.6	-0.5	-0.6	-0.7			
9	m14	0.0	0.3	-0.1	3.7	-0.2	1.9	-0.7	0.4	0.8		
10	m6	0.0	-0.7	0.8	-0.6	-0.7	0.3	0.1	-0.5	2.7	-0.5	
11	m11	0.0	-0.6	2.8	6.9	-0.7	-0.6	0.2	-0.5	-0.1	-0.7	-0.6
12	m7	0.0	-0.6	0.5	0.0	-0.1	-0.4	-0.7	-0.6	-0.7	-0.5	-0.2
13	m5	0.0	0.1	-0.7	0.4	1.3	1.5	-0.6	-0.7	-0.7	4.0	1.1
14	m2	0.0	-0.7	-0.5	1.1	-0.7	1.3	-0.4	1.7	-0.7	1.0	-0.1
15	m10	0.0	2.6	-0.7	0.2	0.0	0.1	0.9	1.1	0.1	2.6	1.8
16	m15	0.0	0.1	-0.7	-0.6	-0.7	-0.6	-0.2	0.9	-0.6	1.8	-0.6
17	m8	0.0	-0.5	-0.5	-0.7	-0.4	0.1	-0.7	-0.6	-0.7	1.4	-0.6
18	m17	0.0	0.2	2.7	-0.0	-0.7	-0.7	-0.7	-0.4	1.0	-0.4	-0.7
19	v3a1	0.0	0.2	1.2	0.6	-0.6	-0.7	-0.1	0.3	-0.5	-0.5	0.5
20	v3a3	0.0	0.3	2.6	-0.5	-0.2	-0.7	-0.2	1.0	-0.4	-0.6	-0.7
21	v3a9	0.0	1.2	-0.7	-0.1	-0.4	-0.7	-0.1	0.4	0.4	1.4	-0.4
22	v3a17	0.1	-0.2	-0.2	-0.4	-0.4	-0.1	1.0	-0.5	-0.4	-0.6	-0.5
23	v3b4	0.0	-0.7	1.0	-0.2	2.5	-0.0	-0.2	-0.6	-0.3	-0.3	-0.6
24	v3b12	0.0	11.4	-0.6	-0.6	-0.3	-0.7	0.4	-0.5	-0.2	-0.5	0.1
25	v3b15	0.0	-0.7	-0.5	-0.6	-0.6	-0.7	-0.3	-0.3	-0.3	1.5	-0.7
26	v3b16	0.0	0.5	-0.7	-0.3	-0.5	-0.7	0.5	4.2	0.0	1.7	-0.6
27	v5a7	0.0	-0.6	1.9	-0.5	-0.4	-0.3	-0.2	0.5	0.8	-0.6	-0.7
28	v5a10	0.0	-0.5	-0.6	0.2	-0.5	-0.4	2.0	-0.2	-0.3	-0.6	-0.7
29	v5a11	0.0	5.1	-0.2	-0.3	1.2	-0.3	-0.2	1.0	-0.3	-0.3	-0.6
30	v5a14	0.0	-0.0	0.8	-0.5	-0.1	-0.6	-0.0	0.1	-0.3	4.1	0.1
31	v5a18	0.0	-0.1	-0.7	-0.2	-0.3	17.1	0.2	-0.5	0.9	-0.7	-0.6
32	v5b2	0.0	-0.5	-0.6	1.0	-0.0	0.1	0.2	-0.5	-0.2	0.6	-0.7
33	v5b5	0.0	-0.6	-0.4	0.2	-0.2	0.8	0.6	0.1	2.5	1.4	-0.7
34	v5b6	0.0	-0.4	-0.5	0.2	2.9	-0.4	0.3	-0.4	0.1	-0.6	2.7
35	v5b8	0.0	0.8	-0.1	-0.0	-0.4	-0.3	1.2	-0.5	0.9	-0.1	-0.2
36	v5b13	0.0	-0.2	-0.6	-0.2	1.8	-0.3	0.7	-0.6	0.9	-0.6	2.2

Item	Label	Marginal X^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.0										
12	m7	0.0	-0.5									
13	m5	0.0	-0.7	0.1								
14	m2	0.0	-0.7	0.4	0.4							
15	m10	0.0	-0.7	-0.6	-0.7	-0.6						
16	m15	0.0	2.5	-0.1	-0.0	1.7	-0.5					
17	m8	0.0	-0.5	-0.7	-0.1	0.6	-0.5	0.3				
18	m17	0.0	-0.6	-0.3	-0.3	-0.5	0.0	-0.7	-0.7			
19	v3a1	0.0	-0.5	-0.7	0.1	0.8	2.4	-0.5	-0.6	0.2		
20	v3a3	0.0	-0.3	-0.5	0.8	-0.5	-0.6	-0.4	0.6	0.6	-0.3	
21	v3a9	0.0	-0.3	-0.4	0.1	-0.5	-0.7	-0.0	-0.4	-0.6	-0.1	-0.5
22	v3a17	0.1	0.7	3.8	-0.1	1.6	-0.4	0.4	0.2	-0.5	1.2	-0.6
23	v3b4	0.0	-0.3	-0.3	-0.2	-0.6	-0.3	0.4	-0.4	-0.4	0.4	-0.7
24	v3b12	0.0	-0.4	-0.4	-0.2	-0.2	-0.6	-0.2	-0.5	0.5	-0.4	-0.7
25	v3b15	0.0	3.1	0.3	2.7	-0.5	-0.5	0.2	-0.0	-0.5	-0.6	-0.7
26	v3b16	0.0	2.3	-0.1	0.2	-0.5	-0.2	0.4	-0.5	-0.6	-0.2	4.2
27	v5a7	0.0	0.5	-0.6	-0.0	-0.7	0.6	0.3	-0.5	-0.2	0.7	-0.4
28	v5a10	0.0	-0.2	-0.0	-0.3	-0.0	4.2	-0.1	-0.6	0.1	-0.0	0.4
29	v5a11	0.0	1.9	-0.0	0.4	-0.4	0.2	1.8	1.3	-0.6	-0.2	-0.2
30	v5a14	0.0	-0.1	-0.6	1.9	-0.5	-0.6	0.2	0.1	-0.7	1.3	3.2
31	v5a18	0.0	0.1	-0.5	0.2	0.1	-0.6	1.7	0.1	-0.7	0.3	0.1
32	v5b2	0.0	2.4	-0.3	-0.4	4.2	-0.4	-0.1	0.3	-0.2	5.9	-0.7
33	v5b5	0.0	-0.1	-0.6	10.1	-0.7	-0.3	-0.1	-0.1	-0.7	-0.3	-0.5
34	v5b6	0.0	0.4	-0.4	1.6	-0.4	-0.2	-0.1	-0.4	-0.6	0.1	-0.5
35	v5b8	0.0	3.7	2.0	-0.3	-0.6	-0.3	-0.1	3.7	2.2	2.2	-0.2
36	v5b13	0.0	-0.3	-0.3	-0.4	-0.2	2.2	-0.1	3.1	-0.5	-0.1	-0.3



		Marginal										
Item	Label	χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.0										
22	v3a17	0.1	-0.6									
23	v3b4	0.0	0.1	-0.3								
24	v3b12	0.0	-0.6	-0.6	0.4							
25	v3b15	0.0	-0.6	-0.7	1.1	-0.6						
26	v3b16	0.0	1.1	-0.7	1.7	-0.6	-0.2					
27	v5a7	0.0	0.2	-0.7	0.5	-0.7	-0.6	-0.4				
28	v5a10	0.0	-0.7	-0.7	1.1	-0.6	-0.3	0.2	-0.1			
29	v5a11	0.0	-0.5	-0.5	-0.7	-0.2	-0.7	1.0	-0.7	-0.7		
30	v5a14	0.0	-0.3	-0.3	0.0	0.3	-0.6	-0.6	-0.5	-0.6	-0.2	
31	v5a18	0.0	-0.5	0.2	0.4	1.1	0.1	-0.6	1.4	-0.5	-0.7	-0.7
32	v5b2	0.0	5.7	0.6	0.9	0.4	-0.5	-0.5	0.7	-0.6	-0.2	-0.6
33	v5b5	0.0	3.3	-0.3	0.2	-0.3	0.1	-0.6	-0.4	-0.6	-0.4	-0.1
34	v5b6	0.0	-0.5	-0.6	-0.0	-0.4	-0.7	-0.5	2.2	-0.4	-0.0	0.9
35	v5b8	0.0	-0.4	2.4	-0.6	-0.6	-0.7	0.5	-0.5	-0.1	0.1	-0.4
36	v5b13	0.0	-0.6	2.5	-0.3	1.0	-0.7	-0.4	-0.4	-0.7	-0.7	1.5



		Marginal					
Item	Label	χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	0.8				
33	v5b5	0.0	-0.7	-0.0			
34	v5b6	0.0	1.4	-0.6	0.1		
35	v5b8	0.0	0.1	3.8	-0.6	-0.4	
36	v5b13	0.0	-0.6	-0.7	-0.6	-0.5	-0.4

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)

-2loglikelihood:	4336.84
Akaike Information Criterion (AIC):	4480.84
Bayesian Information Criterion (BIC):	4676.57

Summary of the Data and Control Parameters

Sample Size	112
Number of Items	36
Number of Dimensions	1

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: VT Sample - all items;
2PL

Item	*Label*		*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e*
1	m12	2	-0.97	0.74	1	1.64	0.66	1.68	0.97
2	m3	4	1.95	0.95	3	-1.08	0.56	0.55	0.29
3	m1	6	0.5	0.45	5	-0.69	0.36	1.37	1.35
4	m13	8	0.28	0.4	7	-0.54	0.35	1.94	3.01
5	m18	10	0.34	0.39	9	0.42	0.35	-1.24	1.73
6	m9	12	0.17	0.39	11	0.17	0.34	-1	2.91
7	m4	14	1.04	0.55	13	0.05	0.4	-0.05	0.39
8	m16	16	2.07	1.04	15	0.64	0.54	-0.31	0.26
9	m14	18	-0.28	0.53	17	1.6	0.46	5.78	10.82
10	m6	20	0.43	0.41	19	-0.18	0.35	0.42	0.89
11	m11	22	0.91	0.52	21	1.23	0.44	-1.36	0.8
12	m7	24	0.6	0.45	23	0.43	0.37	-0.72	0.73
13	m5	26	0.23	0.46	25	-1.19	0.4	5.18	10.45
14	m2	28	-0.33	0.39	27	-0.24	0.35	-0.73	1.31
15	m10	30	0.3	0.42	29	-0.8	0.37	2.64	3.71
16	m15	32	1.11	0.56	31	0.42	0.42	-0.37	0.39
17	m8	34	0.31	0.44	33	1.08	0.4	-3.52	4.98
18	m17	36	0.71	0.47	35	0.18	0.36	-0.26	0.52
19	v3a1	38	0.63	0.45	37	0.7	0.39	-1.12	0.88
20	v3a3	40	0.92	0.55	39	-0.63	0.4	0.68	0.49
21	v3a9	42	-0.27	0.39	41	-0.41	0.35	-1.52	2.44
22	v3a17	44	0.91	0.52	43	0.75	0.41	-0.83	0.56
23	v3b4	46	1.05	0.55	45	0.19	0.4	-0.18	0.4
24	v3b12	48	-1.13	1.14	47	3.31	1.19	2.92	2.23
25	v3b15	50	1.29	0.62	49	0.2	0.42	-0.16	0.33
26	v3b16	52	1.69	0.81	51	-0.81	0.5	0.48	0.31
27	v5a7	54	0.81	0.5	53	-0.47	0.37	0.58	0.55
28	v5a10	56	1.04	0.56	55	-0.09	0.4	0.08	0.38
29	v5a11	58	0.86	0.5	57	-0.34	0.37	0.4	0.47
30	v5a14	60	-0.33	0.47	59	1.25	0.41	3.78	5.29
31	v5a18	62	0.59	0.45	61	0.06	0.35	-0.1	0.6
32	v5b2	64	0.52	0.43	63	0.3	0.35	-0.59	0.77
33	v5b5	66	2.32	1.25	65	-0.57	0.62	0.25	0.24
34	v5b6	68	0.64	0.44	67	0.18	0.35	-0.29	0.59
35	v5b8	70	0.7	0.5	69	0.72	0.39	-1.02	0.8
36	v5b13	72	1.51	0.79	71	1.25	0.59	-0.83	0.38

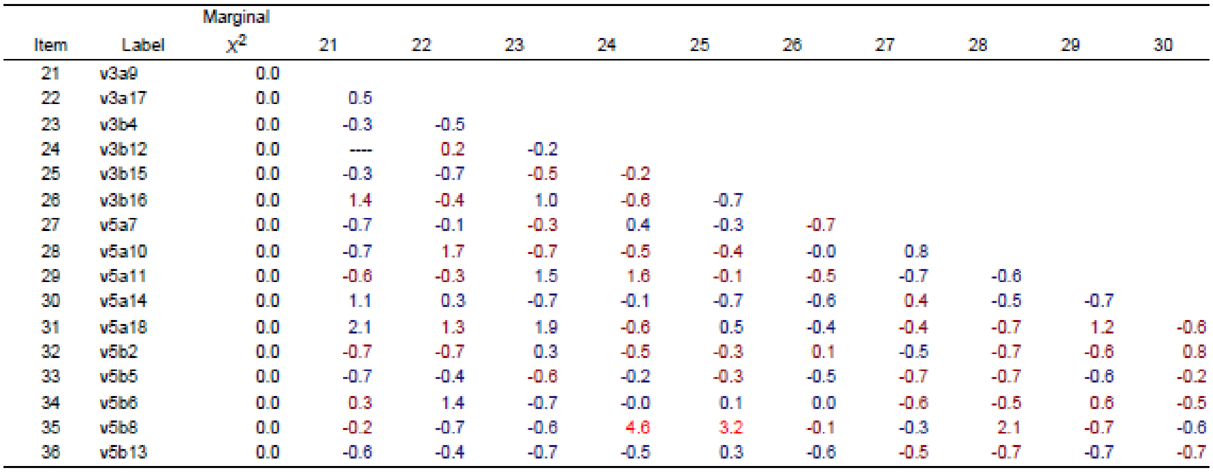
S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	*Probability*
1	m12	3.21	1	0.0729
2	m3	4.38	2	0.1119
3	m1	5.27	2	0.0717
4	m13	3.38	5	0.6419
5	m18	9.8	5	0.0809
6	m9	4.48	5	0.4839
7	m4	3.94	4	0.4155
8	m16	4.66	3	0.1998
9	m14	3.49	1	0.0616
10	m6	8.44	4	0.0767
11	m11	4.85	3	0.1842
12	m7	4.12	3	0.2506
13	m5	3.88	3	0.2761
14	m2	4.69	5	0.4559
15	m10	5.91	3	0.1161
16	m15	7.5	5	0.1855
17	m8	4.67	4	0.3243
18	m17	9.27	4	0.0546
19	v3a1	5.38	4	0.2521
20	v3a3	4.27	3	0.2346
21	v3a9	6.03	5	0.3052
22	v3a17	2.76	4	0.6003
23	v3b4	3.65	5	0.6017
24	v3b12			
25	v3b15	2.41	3	0.493
26	v3b16	4.67	3	0.1988
27	v5a7	9.31	3	0.0254
28	v5a10	5.4	4	0.2501
29	v5a11	2.83	4	0.5873
30	v5a14	3.94	3	0.2697
31	v5a18	9.34	5	0.096
32	v5b2	9.8	4	0.0439
33	v5b5	3.51	2	0.1743
34	v5b6	9.87	6	0.1299
35	v5b8	4.45	4	0.3494
36	v5b13	5.99	3	0.112

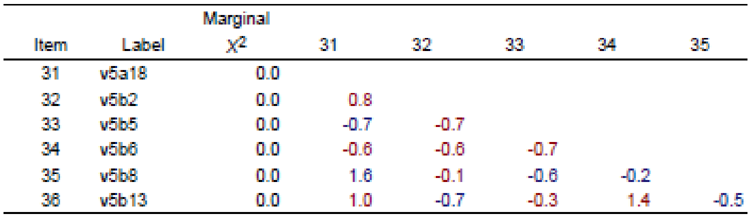
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

		Marginal										
Item	Label	X^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.0										
2	m3	0.0	-0.4									
3	m1	0.0	-0.2	0.6								
4	m13	0.0	-0.7	-0.7	-0.1							
5	m18	0.0	-0.6	-0.7	-0.6	-0.7						
6	m9	0.0	-0.7	2.8	1.3	-0.7	0.1					
7	m4	0.0	-0.5	0.5	-0.3	-0.2	-0.1	-0.4				
8	m16	0.1	-0.5	-0.6	-0.6	-0.2	0.6	-0.5	-0.0			
9	m14	0.0	1.4	-0.5	-0.3	-0.4	-0.7	-0.7	0.4	-0.6		
10	m6	0.0	0.2	-0.6	-0.7	-0.2	-0.4	-0.3	0.6	-0.6	-0.7	
11	m11	0.0	-0.7	2.6	0.4	1.1	-0.3	-0.7	-0.6	-0.5	-0.6	-0.6
12	m7	0.0	-0.3	-0.6	-0.5	-0.4	2.4	-0.7	0.9	-0.4	-0.3	-0.7
13	m5	0.0	-0.2	-0.2	0.1	-0.3	3.1	-0.5	-0.5	-0.7	-0.5	-0.1
14	m2	0.0	0.2	-0.4	-0.3	0.1	0.2	-0.2	-0.5	-0.1	-0.1	0.4
15	m10	0.0	-0.6	-0.6	0.9	-0.7	-0.1	-0.7	-0.7	-0.6	-0.7	-0.7
16	m15	0.0	-0.2	-0.1	-0.6	1.0	0.9	-0.6	-0.2	-0.5	-0.3	-0.7
17	m8	0.0	0.8	-0.6	1.5	-0.0	-0.7	-0.7	-0.4	-0.2	1.1	3.4
18	m17	0.0	0.6	-0.7	0.6	1.6	-0.7	-0.6	2.2	-0.6	-0.5	-0.7
19	v3a1	0.0	0.2	-0.4	-0.6	-0.2	-0.7	-0.7	-0.2	-0.3	0.1	-0.7
20	v3a3	0.0	0.4	-0.5	-0.7	-0.5	-0.7	-0.5	-0.3	-0.3	6.6	-0.6
21	v3a9	0.0	-0.6	-0.4	0.1	-0.4	0.4	0.1	-0.7	-0.6	3.6	-0.5
22	v3a17	0.0	-0.6	-0.6	-0.6	-0.7	-0.7	-0.5	-0.4	-0.2	2.4	-0.2
23	v3b4	0.0	0.1	0.1	-0.5	-0.6	0.6	-0.7	1.6	-0.6	-0.6	-0.6
24	v3b12	0.0	2.0	-0.6	----	----	-0.5	0.3	-0.1	0.8	0.0	0.5
25	v3b15	0.0	-0.7	0.9	-0.6	-0.6	0.7	0.2	-0.7	-0.4	-0.5	-0.6
26	v3b16	0.0	-0.2	-0.5	-0.6	-0.6	1.6	-0.6	1.3	0.2	-0.5	-0.7
27	v5a7	0.0	-0.5	-0.6	-0.6	1.4	-0.4	-0.7	-0.6	-0.5	-0.6	-0.4
28	v5a10	0.0	0.6	-0.4	-0.6	-0.7	-0.7	0.1	-0.1	-0.6	-0.7	0.1
29	v5a11	0.0	3.6	-0.4	-0.7	-0.7	-0.6	-0.0	-0.3	-0.2	-0.7	0.0
30	v5a14	0.0	0.1	-0.5	-0.7	-0.1	-0.7	0.4	-0.6	-0.5	-0.5	-0.1
31	v5a18	0.0	1.9	-0.6	-0.1	0.7	0.4	-0.7	-0.2	-0.3	2.2	0.3
32	v5b2	0.0	0.6	-0.3	-0.5	-0.1	-0.6	-0.4	-0.5	-0.6	-0.5	-0.6
33	v5b5	0.0	0.4	0.2	-0.7	-0.7	-0.4	1.6	-0.7	-0.5	-0.7	-0.4
34	v5b6	0.0	-0.6	-0.7	-0.3	-0.6	-0.7	-0.6	0.4	-0.1	0.7	-0.7
35	v5b8	0.0	-0.7	-0.4	-0.6	-0.4	-0.7	-0.5	-0.4	3.7	-0.7	-0.3
36	v5b13	0.0	-0.7	-0.6	0.2	-0.1	0.1	-0.2	-0.6	-0.6	-0.6	1.3

		Marginal										
Item	Label	X^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.0										
12	m7	0.0	-0.7									
13	m5	0.0	-0.6	4.0								
14	m2	0.0	0.4	-0.7	-0.6							
15	m10	0.0	-0.6	-0.4	0.4	-0.7						
16	m15	0.0	-0.1	2.1	-0.6	-0.6	0.5					
17	m8	0.0	-0.5	1.3	-0.0	1.3	-0.7	-0.7				
18	m17	0.0	-0.2	-0.5	-0.6	0.1	0.4	0.4	-0.6			
19	v3a1	0.0	-0.4	-0.6	-0.5	-0.6	-0.6	-0.3	0.5	-0.2		
20	v3a3	0.0	0.3	-0.0	-0.5	-0.2	0.2	-0.7	-0.7	-0.7	0.1	
21	v3a9	0.0	1.1	0.8	0.0	0.9	-0.7	-0.7	-0.7	-0.2	-0.4	1.1
22	v3a17	0.0	-0.7	-0.6	-0.5	-0.6	0.1	2.1	0.4	0.1	0.9	-0.5
23	v3b4	0.0	-0.7	0.2	0.3	-0.2	-0.6	-0.5	0.3	0.0	-0.6	0.2
24	v3b12	0.0	-0.5	-0.2	----	----	----	-0.0	0.1	-0.1	-0.1	0.4
25	v3b15	0.0	0.4	0.0	-0.6	-0.6	0.1	-0.6	0.1	-0.6	-0.6	0.3
26	v3b16	0.0	-0.1	-0.3	-0.0	0.1	-0.7	-0.6	0.9	0.2	-0.7	-0.6
27	v5a7	0.0	0.1	0.5	0.7	-0.5	2.3	1.7	-0.7	3.4	2.7	-0.5
28	v5a10	0.0	-0.7	1.0	1.3	-0.7	-0.2	1.4	-0.2	-0.0	-0.7	-0.5
29	v5a11	0.0	1.0	-0.7	0.3	0.3	-0.2	0.1	0.3	-0.5	-0.7	-0.2
30	v5a14	0.0	-0.7	1.4	-0.7	-0.5	-0.6	3.5	-0.7	5.6	-0.6	-0.5
31	v5a18	0.0	-0.7	-0.6	-0.7	1.6	-0.4	-0.3	-0.7	1.3	0.2	0.2
32	v5b2	0.0	1.9	0.0	-0.6	0.1	-0.6	-0.3	0.8	-0.6	3.3	1.0
33	v5b5	0.0	0.1	-0.6	-0.2	-0.5	-0.6	-0.6	0.7	-0.3	-0.7	-0.5
34	v5b6	0.0	-0.7	-0.3	-0.6	-0.6	1.3	0.4	0.1	0.5	-0.4	-0.3
35	v5b8	0.0	-0.4	0.1	-0.6	-0.7	-0.6	-0.5	0.5	-0.1	-0.5	0.2
36	v5b13	0.0	-0.7	-0.7	1.6	-0.3	0.2	-0.6	-0.7	-0.2	-0.6	-0.7



		Marginal										
Item	Label	χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.0										
22	v3a17	0.0	0.5									
23	v3b4	0.0	-0.3	-0.5								
24	v3b12	0.0	----	0.2	-0.2							
25	v3b15	0.0	-0.3	-0.7	-0.5	-0.2						
26	v3b16	0.0	1.4	-0.4	1.0	-0.6	-0.7					
27	v5a7	0.0	-0.7	-0.1	-0.3	0.4	-0.3	-0.7				
28	v5a10	0.0	-0.7	1.7	-0.7	-0.5	-0.4	-0.0	0.8			
29	v5a11	0.0	-0.6	-0.3	1.5	1.6	-0.1	-0.5	-0.7	-0.6		
30	v5a14	0.0	1.1	0.3	-0.7	-0.1	-0.7	-0.6	0.4	-0.5	-0.7	
31	v5a18	0.0	2.1	1.3	1.9	-0.6	0.5	-0.4	-0.4	-0.7	1.2	-0.6
32	v5b2	0.0	-0.7	-0.7	0.3	-0.5	-0.3	0.1	-0.5	-0.7	-0.6	0.8
33	v5b5	0.0	-0.7	-0.4	-0.6	-0.2	-0.3	-0.5	-0.7	-0.7	-0.6	-0.2
34	v5b6	0.0	0.3	1.4	-0.7	-0.0	0.1	0.0	-0.6	-0.5	0.6	-0.5
35	v5b8	0.0	-0.2	-0.7	-0.6	4.6	3.2	-0.1	-0.3	2.1	-0.7	-0.6
36	v5b13	0.0	-0.6	-0.4	-0.7	-0.5	0.3	-0.6	-0.5	-0.7	-0.7	-0.7



		Marginal					
Item	Label	χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	0.8				
33	v5b5	0.0	-0.7	-0.7			
34	v5b6	0.0	-0.6	-0.6	-0.7		
35	v5b8	0.0	1.6	-0.1	-0.6	-0.2	
36	v5b13	0.0	1.0	-0.7	-0.3	1.4	-0.5

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	1489.64
Akaike Information Criterion (AIC):	1633.64
Bayesian Information Criterion (BIC):	1745.63

Summary of the Data and Control Parameters

Sample Size	35
Number of Items	36
Number of Dimensions	1

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: National Sample - all items; 3PL

Item	*Label*		*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e*		*logit g*	*s.e.*	*g*	*s.e.*
1	m12	3	0.12	0.08	2	-1.09	0.59	8.93	7.45	1	-1.16	0.52	0.24	0.09
2	m3	6	0.95	0.32	5	0.04	0.35	-0.05	0.38	4	-1.46	0.55	0.19	0.08
3	m1	9	0.87	0.32	8	-0.36	0.41	0.42	0.42	7	-1.43	0.54	0.19	0.08
4	m13	12	0.5	0.22	11	0.19	0.32	-0.38	0.68	10	-1.4	0.56	0.2	0.09
5	m18	15	0.19	0.12	14	-0.66	0.43	3.55	3.17	13	-1.3	0.54	0.21	0.09
6	m9	18	1.43	0.47	17	-0.92	0.49	0.64	0.25	16	-1.86	0.56	0.13	0.06
7	m4	21	1.93	0.54	20	2.11	0.47	-1.09	0.25	19	-1.56	0.57	0.17	0.08
8	m16	24	1.28	0.37	23	0.46	0.34	-0.36	0.3	22	-1.53	0.56	0.18	0.08
9	m14	27	0.61	0.28	26	1.75	0.35	-2.86	1.27	25	-1.4	0.56	0.2	0.09
10	m6	30	0.55	0.25	29	-0.37	0.39	0.67	0.66	28	-1.4	0.55	0.2	0.09
11	m11	33	0.19	0.12	32	1.01	0.29	-5.31	3.76	31	-1.37	0.56	0.2	0.09
12	m7	36	1.32	0.42	35	2.04	0.43	-1.54	0.42	34	-1.4	0.56	0.2	0.09
13	m5	39	0.52	0.26	38	-0.58	0.42	1.11	0.77	37	-1.42	0.55	0.2	0.09
14	m2	42	0.23	0.15	41	-0.73	0.45	3.2	2.7	40	-1.28	0.53	0.22	0.09
15	m10	45	0.25	0.16	44	-0.65	0.43	2.62	2.2	43	-1.34	0.54	0.21	0.09
16	m15	48	0.59	0.28	47	-0.55	0.43	0.92	0.66	46	-1.33	0.54	0.21	0.09
17	m8	51	1.2	0.36	50	1.27	0.35	-1.06	0.37	49	-1.4	0.55	0.2	0.09
18	m17	54	0.88	0.29	53	0.34	0.33	-0.39	0.42	52	-1.43	0.56	0.19	0.09
19	v3a1	57	0.75	0.28	56	0.3	0.34	-0.4	0.5	55	-1.36	0.55	0.2	0.09
20	v3a3	60	0.96	0.42	59	-0.96	0.58	1.01	0.45	58	-1.4	0.51	0.2	0.08
21	v3a9	63	0.73	0.35	62	-1.02	0.54	1.39	0.64	61	-1.58	0.56	0.17	0.08
22	v3a17	66	1.48	0.45	65	1.43	0.39	-0.96	0.31	64	-1.39	0.55	0.2	0.09
23	v3b4	69	1.42	0.44	68	0.98	0.37	-0.68	0.32	67	-1.37	0.54	0.2	0.09
24	v3b12	72	0.31	0.19	71	0.94	0.31	-3	2.03	70	-1.39	0.56	0.2	0.09
25	v3b15	75	0.25	0.22	74	-17.06	33E+5	67.08	13E+6	73	-0.79	0.42	0.31	0.09
26	v3b16	78	0.79	0.35	77	-0.56	0.47	0.71	0.51	76	-1.33	0.53	0.21	0.09
27	v5a7	81	1.49	0.49	80	0.07	0.42	-0.04	0.29	79	-1.43	0.52	0.19	0.08
28	v5a10	84	0.72	0.3	83	-0.33	0.41	0.47	0.53	82	-1.44	0.56	0.19	0.09
29	v5a11	87	0.49	0.27	86	-0.65	0.47	1.32	0.95	85	-1.33	0.54	0.21	0.09
30	v5a14	90	0.62	0.28	89	1.19	0.34	-1.9	0.91	88	-1.42	0.56	0.19	0.09
31	v5a18	93	0.37	0.21	92	-0.34	0.39	0.92	1.08	91	-1.4	0.56	0.2	0.09
32	v5b2	96	0.56	0.27	95	-0.2	0.39	0.36	0.66	94	-1.41	0.56	0.2	0.09
33	v5b5	99	0.23	0.15	98	0.65	0.32	-2.83	2.28	97	-1.37	0.56	0.2	0.09
34	v5b6	102	1.09	0.38	101	0.11	0.39	-0.1	0.37	100	-1.45	0.55	0.19	0.08
35	v5b8	105	1.58	0.5	104	1.26	0.41	-0.8	0.3	103	-1.37	0.54	0.2	0.09
36	v5b13	108	1.71	0.56	107	2.15	0.54	-1.25	0.32	106	-1.42	0.56	0.19	0.09

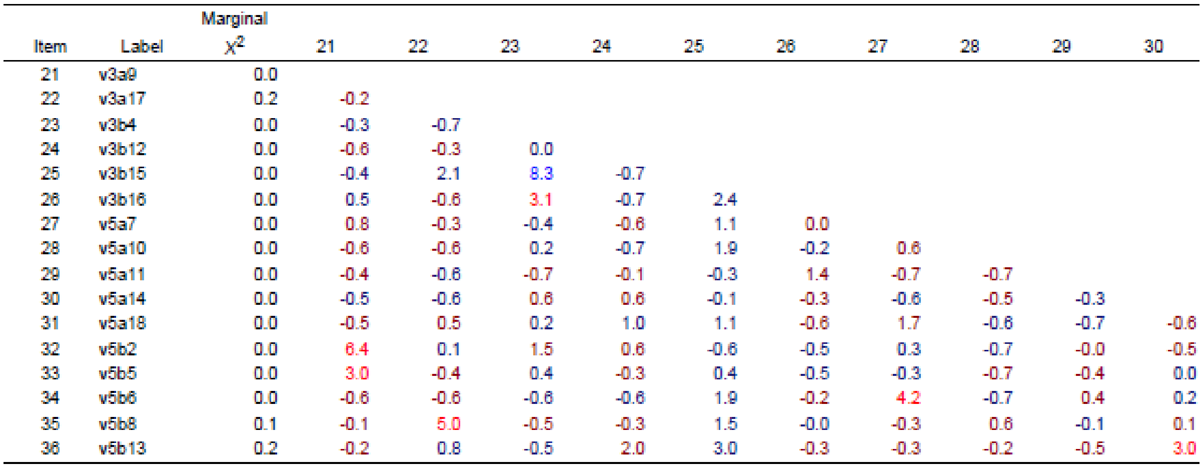
S-X^2 Item Level Diagnostic Statistics

Item	**Label**	X^2	***d.f.***	***Probability***
1	m12	16.38	11	0.1271
2	m3	14.9	11	0.1868
3	m1	9.09	11	0.6142
4	m13	11.47	9	0.2443
5	m18	16.32	13	0.2316
6	m9	8.37	11	0.6805
7	m4	8.45	4	0.0761
8	m16	18.56	9	0.0292
9	m14	11.02	5	0.0508
10	m6	10.49	11	0.4883
11	m11	10.96	8	0.2033
12	m7	6.81	2	0.0332
13	m5	9.83	9	0.3658
14	m2	14.17	12	0.289
15	m10	16.52	11	0.1225
16	m15	7.4	11	0.7669
17	m8	10.26	6	0.1138
18	m17	10.19	7	0.1776
19	v3a1	9.36	8	0.315
20	v3a3	14.54	12	0.2667
21	v3a9	12.2	9	0.2018
22	v3a17	14.3	5	0.0138
23	v3b4	7.04	5	0.2173
24	v3b12	16.57	8	0.0348
25	v3b15	28.03	10	0.0018
26	v3b16	6.3	12	0.9007
27	v5a7	10.42	8	0.2364
28	v5a10	17.25	9	0.0449
29	v5a11	16.05	12	0.1883
30	v5a14	8.35	6	0.213
31	v5a18	11.51	13	0.5694
32	v5b2	8.83	9	0.4541
33	v5b5	9.14	10	0.5203
34	v5b6	10.13	9	0.3414
35	v5b8	8.79	6	0.1852
36	v5b13	4.19	3	0.2431

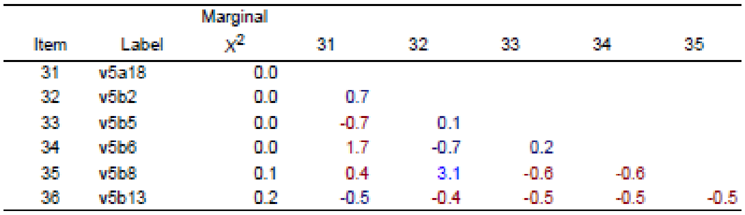
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

		Marginal										
Item	Label	X^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.0										
2	m3	0.0	1.6									
3	m1	0.0	-0.5	-0.2								
4	m13	0.0	2.4	0.9	-0.1							
5	m18	0.0	1.5	-0.6	-0.6	0.6						
6	m9	0.0	-0.6	-0.7	3.4	-0.6	-0.1					
7	m4	0.2	0.3	1.2	0.1	-0.3	-0.2	-0.6				
8	m16	0.0	2.6	-0.3	0.1	-0.4	0.2	-0.6	-0.2			
9	m14	0.0	2.2	-0.4	3.0	0.1	2.8	-0.5	1.6	1.8		
10	m6	0.0	-0.1	0.2	-0.4	-0.7	0.8	-0.3	0.3	4.2	-0.3	
11	m11	0.0	-0.6	1.6	5.3	-0.5	-0.6	-0.4	-0.6	0.8	-0.6	-0.7
12	m7	0.0	0.4	1.8	0.8	-0.4	-0.7	-0.3	-0.4	-0.1	-0.7	-0.5
13	m5	0.0	-0.6	-0.7	0.8	0.9	0.9	-0.4	-0.5	-0.5	3.4	1.6
14	m2	0.0	-0.6	-0.2	0.5	-0.7	1.2	-0.7	3.6	-0.5	1.5	0.2
15	m10	0.0	2.2	-0.6	-0.0	-0.1	-0.0	0.5	2.0	0.4	3.0	2.1
16	m15	0.0	-0.6	-0.6	-0.7	-0.6	-0.7	0.1	0.5	-0.7	1.4	-0.4
17	m8	0.0	2.2	-0.7	-0.6	-0.6	1.1	-0.4	-0.5	-0.2	2.6	-0.3
18	m17	0.0	-0.7	1.5	-0.3	-0.7	-0.4	-0.6	0.5	2.4	-0.1	-0.7
19	v3a1	0.0	2.6	0.5	1.0	-0.6	-0.6	-0.1	0.2	-0.5	-0.6	0.1
20	v3a3	0.0	3.0	3.9	-0.4	-0.5	-0.5	-0.0	0.3	0.3	-0.7	-0.6
21	v3a9	0.0	-0.2	-0.6	0.1	-0.5	-0.5	-0.0	1.8	-0.1	1.0	-0.6
22	v3a17	0.2	4.0	-0.6	-0.4	-0.5	-0.5	2.5	1.1	-0.3	-0.4	-0.5
23	v3b4	0.0	0.7	0.1	-0.5	1.7	0.9	-0.1	-0.2	0.7	0.1	-0.7
24	v3b12	0.0	14.6	-0.7	-0.6	-0.4	-0.7	0.2	-0.4	-0.3	-0.4	-0.1
25	v3b15	0.0	-0.2	0.0	-0.2	-0.2	-0.6	0.8	1.9	0.3	0.1	-0.2
26	v3b16	0.0	3.2	-0.7	-0.5	-0.6	-0.6	0.1	2.5	1.0	1.1	-0.4
27	v5a7	0.0	0.5	3.3	-0.4	-0.2	-0.5	0.1	-0.1	-0.0	-0.5	-0.7
28	v5a10	0.0	-0.5	-0.6	0.7	-0.4	-0.0	3.1	-0.1	-0.0	-0.7	-0.6
29	v5a11	0.0	7.8	0.0	-0.4	1.5	-0.1	-0.1	1.0	-0.2	-0.2	-0.7
30	v5a14	0.0	-0.6	0.3	-0.4	0.1	-0.5	-0.1	1.3	0.0	5.2	-0.2
31	v5a18	0.0	-0.5	-0.6	-0.3	-0.2	16.0	0.4	0.0	0.7	-0.7	-0.6
32	v5b2	0.0	0.2	-0.6	1.2	-0.1	0.5	0.4	0.1	0.0	0.3	-0.6
33	v5b5	0.0	-0.5	-0.3	0.1	-0.2	0.7	0.8	0.3	3.0	1.6	-0.7
34	v5b6	0.0	2.1	-0.7	-0.1	4.1	-0.1	0.4	-0.1	1.4	-0.3	4.0
35	v5b8	0.1	-0.5	-0.5	-0.2	-0.3	-0.3	2.2	0.5	0.1	0.5	-0.4
36	v5b13	0.2	0.2	0.0	0.4	3.1	0.6	1.8	0.5	0.1	-0.5	1.3

		Marginal										
Item	Label	X^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.0										
12	m7	0.0	-0.7									
13	m5	0.0	-0.7	-0.3								
14	m2	0.0	-0.6	1.2	0.1							
15	m10	0.0	-0.7	-0.7	-0.6	-0.7						
16	m15	0.0	3.2	-0.4	0.2	1.2	-0.6					
17	m8	0.0	0.1	-0.4	-0.4	1.4	-0.2	-0.1				
18	m17	0.0	-0.7	0.4	0.1	-0.2	0.3	-0.7	-0.6			
19	v3a1	0.0	-0.4	-0.4	-0.1	0.4	2.0	-0.5	-0.2	0.7		
20	v3a3	0.0	0.1	-0.7	1.3	-0.5	-0.5	-0.2	-0.1	-0.1	-0.6	
21	v3a9	0.0	-0.4	0.2	0.2	-0.5	-0.7	0.2	-0.5	-0.4	0.2	-0.6
22	v3a17	0.2	1.8	1.6	0.4	2.9	-0.1	0.1	1.9	-0.6	0.4	0.1
23	v3b4	0.0	0.3	-0.5	-0.0	-0.3	-0.1	0.9	0.4	0.2	1.2	-0.4
24	v3b12	0.0	-0.4	-0.6	-0.2	-0.3	-0.6	-0.1	-0.6	0.2	-0.3	-0.7
25	v3b15	0.0	3.6	4.7	5.7	-0.4	-0.4	1.7	-0.3	-0.2	0.1	0.1
26	v3b16	0.0	3.4	-0.6	-0.1	-0.6	-0.1	0.7	-0.7	-0.7	0.1	5.6
27	v5a7	0.0	1.3	-0.4	-0.3	-0.5	0.3	0.7	-0.6	-0.6	0.3	0.1
28	v5a10	0.0	-0.2	0.5	-0.1	0.3	3.8	-0.1	-0.6	-0.3	-0.2	-0.2
29	v5a11	0.0	1.4	-0.4	0.2	-0.5	0.4	1.6	0.8	-0.7	-0.1	0.0
30	v5a14	0.0	0.1	-0.5	1.4	-0.6	-0.6	0.4	-0.3	-0.7	1.8	2.3
31	v5a18	0.0	-0.0	-0.6	0.1	-0.0	-0.5	1.7	0.2	-0.7	0.3	-0.1
32	v5b2	0.0	1.7	-0.0	-0.3	3.6	-0.4	-0.1	-0.1	0.1	6.5	-0.7
33	v5b5	0.0	-0.2	-0.6	9.8	-0.6	-0.3	0.0	0.0	-0.7	-0.2	-0.5
34	v5b6	0.0	1.1	-0.6	0.8	-0.6	0.0	0.1	-0.7	-0.6	-0.2	0.1
35	v5b8	0.1	5.3	0.7	0.1	-0.4	-0.3	0.1	5.9	1.2	1.5	-0.6
36	v5b13	0.2	-0.3	0.8	-0.4	-0.3	2.0	0.3	1.6	0.2	0.5	-0.5



Item	Label	Marginal χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.0										
22	v3a17	0.2	-0.2									
23	v3b4	0.0	-0.3	-0.7								
24	v3b12	0.0	-0.6	-0.3	0.0							
25	v3b15	0.0	-0.4	2.1	8.3	-0.7						
26	v3b16	0.0	0.5	-0.6	3.1	-0.7	2.4					
27	v5a7	0.0	0.8	-0.3	-0.4	-0.6	1.1	0.0				
28	v5a10	0.0	-0.6	-0.6	0.2	-0.7	1.9	-0.2	0.6			
29	v5a11	0.0	-0.4	-0.6	-0.7	-0.1	-0.3	1.4	-0.7	-0.7		
30	v5a14	0.0	-0.5	-0.6	0.6	0.6	-0.1	-0.3	-0.6	-0.5	-0.3	
31	v5a18	0.0	-0.5	0.5	0.2	1.0	1.1	-0.6	1.7	-0.6	-0.7	-0.6
32	v5b2	0.0	6.4	0.1	1.5	0.6	-0.6	-0.5	0.3	-0.7	-0.0	-0.5
33	v5b5	0.0	3.0	-0.4	0.4	-0.3	0.4	-0.5	-0.3	-0.7	-0.4	0.0
34	v5b6	0.0	-0.6	-0.6	-0.6	-0.6	1.9	-0.2	4.2	-0.7	0.4	0.2
35	v5b8	0.1	-0.1	5.0	-0.5	-0.3	1.5	-0.0	-0.3	0.6	-0.1	0.1
36	v5b13	0.2	-0.2	0.8	-0.5	2.0	3.0	-0.3	-0.3	-0.2	-0.5	3.0



Item	Label	Marginal χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	0.7				
33	v5b5	0.0	-0.7	0.1			
34	v5b6	0.0	1.7	-0.7	0.2		
35	v5b8	0.1	0.4	3.1	-0.6	-0.6	
36	v5b13	0.2	-0.5	-0.4	-0.5	-0.5	-0.5

Likelihood-based Values and Goodness of Fit Statistics	
Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	4362.18
Akaike Information Criterion (AIC):	4578.18
Bayesian Information Criterion (BIC):	4871.78

Summary of the Data and Control Parameters	
Sample Size	112
Number of Items	36
Number of Dimensions	1

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: VT Sample - all items; 3PL

Item	***Label***		***a***	***s.e.***		***c***	***s.e.***	***b***	***s.e***		***logit g***	***s.e.***	***g***	***s.e.***
1	m12	3	0.19	0.14	2	1.1	0.48	-5.69	4.86	1	-1.37	0.56	0.2	0.09
2	m3	6	1.47	1.02	5	-1.52	1.13	1.03	0.54	4	-1.77	0.61	0.15	0.08
3	m1	9	0.7	0.71	8	-1.63	1.2	2.31	1.63	7	-1.44	0.58	0.19	0.09
4	m13	12	0.47	0.43	11	-1.37	0.87	2.89	2.41	10	-1.4	0.56	0.2	0.09
5	m18	15	0.38	0.29	14	0.01	0.49	-0.01	1.3	13	-1.39	0.55	0.2	0.09
6	m9	18	0.28	0.21	17	-0.32	0.52	1.16	2.01	16	-1.36	0.55	0.2	0.09
7	m4	21	1.66	1.31	20	-0.81	1.01	0.49	0.48	19	-1.31	0.53	0.21	0.09
8	m16	24	1.48	0.98	23	0.28	0.67	-0.19	0.44	22	-1.51	0.56	0.18	0.08
9	m14	27	0.28	0.22	26	1.33	0.49	-4.66	3.89	25	-1.38	0.57	0.2	0.09
10	m6	30	0.31	0.25	29	-0.79	0.62	2.55	2.62	28	-1.37	0.56	0.2	0.09
11	m11	33	0.38	0.3	32	0.79	0.47	-2.07	1.97	31	-1.4	0.57	0.2	0.09
12	m7	36	0.6	0.42	35	0.01	0.53	-0.02	0.88	34	-1.39	0.57	0.2	0.09
13	m5	39	0.43	0.45	38	-2.67	1.93	6.21	6.92	37	-1.47	0.53	0.19	0.08
14	m2	42	0.22	0.17	41	-0.98	0.7	4.37	4.41	40	-1.29	0.54	0.22	0.09
15	m10	45	0.39	0.36	44	-1.77	1.02	4.51	4.33	43	-1.43	0.54	0.19	0.08
16	m15	48	0.63	0.44	47	0.02	0.52	-0.03	0.81	46	-1.48	0.56	0.19	0.08
17	m8	51	0.58	0.41	50	0.84	0.49	-1.45	1.26	49	-1.39	0.55	0.2	0.09
18	m17	54	0.47	0.35	53	-0.28	0.52	0.59	1.12	52	-1.41	0.57	0.2	0.09
19	v3a1	57	0.67	0.45	56	0.39	0.51	-0.58	0.86	55	-1.44	0.59	0.19	0.09
20	v3a3	60	0.46	0.4	59	-1.27	0.8	2.73	2.36	58	-1.46	0.58	0.19	0.09
21	v3a9	63	0.22	0.18	62	-1.29	0.87	5.8	5.76	61	-1.27	0.57	0.22	0.1
22	v3a17	66	0.93	0.6	65	0.39	0.55	-0.42	0.62	64	-1.4	0.57	0.2	0.09
23	v3b4	69	0.77	0.55	68	-0.34	0.64	0.45	0.75	67	-1.38	0.57	0.2	0.09
24	v3b12	72	0.26	0.21	71	2.6	0.75	-10.16	8.59	70	-1.38	0.56	0.2	0.09
25	v3b15	75	0.98	0.63	74	-0.19	0.59	0.2	0.59	73	-1.53	0.57	0.18	0.08
26	v3b16	78	1.51	1.04	77	-1.28	0.93	0.85	0.54	76	-1.78	0.58	0.14	0.07
27	v5a7	81	0.79	0.75	80	-1.25	1.08	1.59	1.06	79	-1.41	0.59	0.2	0.09
28	v5a10	84	0.73	0.51	83	-0.55	0.6	0.74	0.84	82	-1.5	0.56	0.18	0.08
29	v5a11	87	0.73	0.54	86	-0.88	0.73	1.21	0.94	85	-1.52	0.6	0.18	0.09
30	v5a14	90	0.21	0.16	89	0.92	0.47	-4.38	3.97	88	-1.37	0.56	0.2	0.09
31	v5a18	93	0.6	0.43	92	-0.43	0.56	0.71	0.95	91	-1.44	0.58	0.19	0.09
32	v5b2	96	0.84	0.54	95	-0.11	0.56	0.13	0.66	94	-1.44	0.58	0.19	0.09
33	v5b5	99	1.04	0.71	98	-0.85	0.81	0.82	0.63	97	-1.62	0.63	0.16	0.09
34	v5b6	102	0.35	0.27	101	-0.27	0.51	0.79	1.53	100	-1.41	0.58	0.2	0.09
35	v5b8	105	0.82	0.56	104	0.39	0.53	-0.47	0.67	103	-1.42	0.55	0.19	0.09
36	v5b13	108	0.91	0.57	107	0.8	0.54	-0.88	0.72	106	-1.45	0.58	0.19	0.09

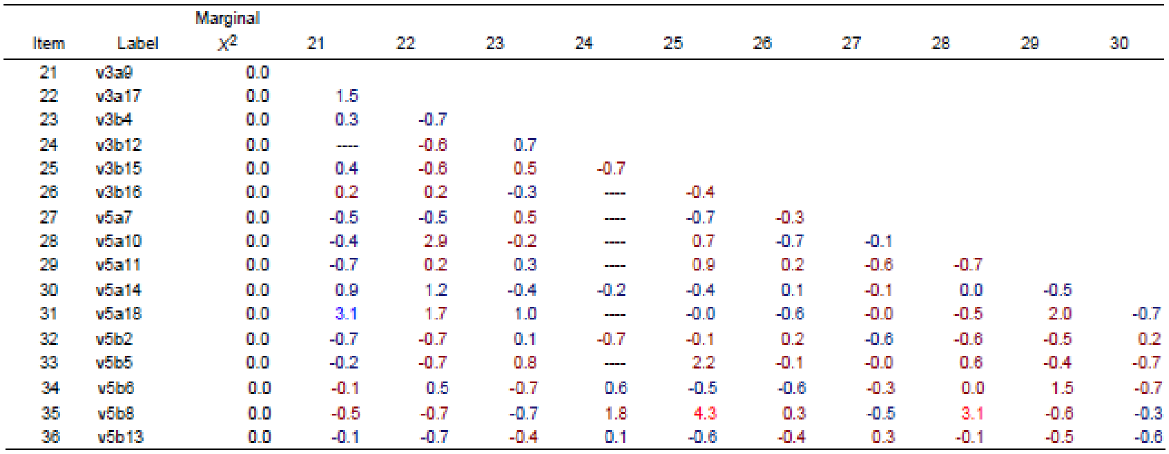
S-X^2 Item Level Diagnostic Statistics

Item	**Label**	X^2	***d.f.***	***Probability***
1	m12			
2	m3	7.88	3	0.0484
3	m1	5.24	1	0.022
4	m13	3.31	4	0.5082
5	m18	12.12	4	0.0165
6	m9	4.55	4	0.3383
7	m4	2.9	2	0.2356
8	m16	6.84	3	0.0771
9	m14	3.33	1	0.0681
10	m6	7	4	0.1354
11	m11	5.75	1	0.0165
12	m7	4.08	2	0.1301
13	m5	3.98	2	0.1381
14	m2	10.03	5	0.0741
15	m10	7.87	3	0.0487
16	m15	7.93	4	0.0941
17	m8	4.19	3	0.2429
18	m17	10.46	4	0.0333
19	v3a1	5.7	3	0.1272
20	v3a3	8.18	4	0.085
21	v3a9	5.47	5	0.3621
22	v3a17	2.84	3	0.4179
23	v3b4	5.04	3	0.1683
24	v3b12			
25	v3b15	4.63	3	0.2023
26	v3b16	6.63	4	0.1564
27	v5a7	8.32	3	0.0397
28	v5a10	3.78	5	0.5827
29	v5a11	3.39	4	0.4959
30	v5a14	5	3	0.1734
31	v5a18	9.2	4	0.0561
32	v5b2	8.65	3	0.0343
33	v5b5	5.94	2	0.0511
34	v5b6	9.37	5	0.0949
35	v5b8	4.26	3	0.2363
36	v5b13	10.09	3	0.0178

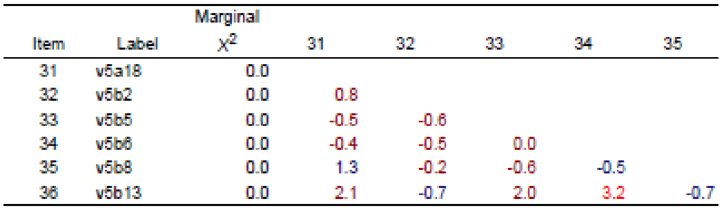
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

Item	Label	Marginal χ^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.0										
2	m3	0.0	-0.3									
3	m1	0.0	-0.6	1.4								
4	m13	0.0	-0.6	-0.7	0.0							
5	m18	0.0	-0.3	-0.6	-0.6	-0.7						
6	m9	0.0	-0.7	2.9	1.3	-0.7	0.1					
7	m4	0.0	-0.5	1.6	-0.2	-0.2	-0.1	-0.5				
8	m16	0.0	1.3	0.1	-0.7	-0.4	1.0	-0.6	-0.4			
9	m14	0.0	2.0	0.2	0.0	-0.3	-0.6	-0.7	-0.3	-0.5		
10	m6	0.0	1.0	-0.7	-0.7	-0.3	-0.5	-0.2	1.2	-0.5	-0.6	
11	m11	0.0	-0.2	0.1	-0.1	0.8	-0.4	-0.7	-0.7	-0.5	-0.4	-0.3
12	m7	0.0	-0.7	-0.2	-0.6	-0.4	2.2	-0.7	1.3	-0.7	-0.6	-0.7
13	m5	0.0	-0.4	0.2	0.2	-0.3	2.9	-0.4	-0.4	-0.6	-0.6	-0.0
14	m2	0.0	-0.2	-0.6	-0.5	-0.2	0.6	-0.3	-0.6	1.5	-0.2	0.9
15	m10	0.0	-0.7	-0.3	1.1	-0.7	-0.0	-0.7	-0.7	-0.7	-0.7	-0.7
16	m15	0.0	1.4	-0.7	-0.6	0.7	1.3	-0.6	-0.6	0.1	-0.6	-0.6
17	m8	0.0	0.1	-0.5	1.6	-0.1	-0.7	-0.7	-0.2	-0.3	0.6	3.4
18	m17	0.0	-0.5	-0.4	1.1	1.9	-0.7	-0.6	1.2	-0.5	-0.3	-0.7
19	v3a1	0.0	-0.6	0.1	-0.7	-0.1	-0.7	-0.7	-0.1	-0.6	-0.3	-0.7
20	v3a3	0.0	2.6	1.0	-0.7	-0.3	-0.7	-0.5	-0.7	-0.6	8.5	-0.7
21	v3a9	0.0	-0.7	-0.7	-0.2	-0.5	0.1	0.2	-0.6	-0.3	3.5	-0.6
22	v3a17	0.0	-0.5	-0.7	-0.5	-0.7	-0.7	-0.5	-0.3	-0.6	1.2	-0.5
23	v3b4	0.0	2.0	1.9	-0.3	-0.7	0.3	-0.6	2.8	-0.1	-0.2	-0.3
24	v3b12	0.0	5.1	----	----	----	-0.7	0.6	----	-0.7	0.3	----
25	v3b15	0.0	0.3	-0.6	-0.7	-0.6	0.4	0.2	-0.4	0.8	-0.7	-0.4
26	v3b16	0.0	-0.5	1.2	-0.7	-0.6	2.0	-0.6	2.2	2.0	-0.7	-0.5
27	v5a7	0.0	-0.7	0.1	-0.4	1.7	-0.5	-0.7	-0.7	0.3	-0.4	-0.0
28	v5a10	0.0	3.1	-0.6	-0.7	-0.6	-0.6	-0.0	-0.5	0.2	-0.7	-0.4
29	v5a11	0.0	1.1	-0.7	-0.6	-0.7	-0.7	-0.1	-0.6	-0.7	-0.5	-0.3
30	v5a14	0.0	0.6	0.4	-0.6	-0.3	-0.7	0.7	-0.1	-0.6	-0.6	0.2
31	v5a18	0.0	0.4	-0.7	0.1	0.8	0.5	-0.7	-0.3	-0.6	1.5	-0.0
32	v5b2	0.0	-0.3	-0.1	-0.5	-0.1	-0.6	-0.3	-0.6	-0.7	-0.7	-0.6
33	v5b5	0.0	4.5	-0.4	-0.6	-0.7	-0.6	1.3	-0.3	0.2	-0.5	0.4
34	v5b6	0.0	0.1	-0.3	-0.5	-0.5	-0.6	-0.6	-0.2	-0.6	1.2	-0.7
35	v5b8	0.0	-0.5	0.0	-0.7	-0.4	-0.7	-0.5	-0.4	4.9	-0.7	-0.5
36	v5b13	0.0	0.1	0.6	-0.2	0.2	0.4	-0.1	-0.7	-0.1	-0.1	0.5

Item	Label	Marginal χ^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.0										
12	m7	0.0	-0.7									
13	m5	0.0	-0.6	4.2								
14	m2	0.0	1.3	-0.6	-0.6							
15	m10	0.0	-0.4	-0.5	0.3	-0.7						
16	m15	0.0	1.2	1.1	-0.5	-0.6	0.9					
17	m8	0.0	-0.5	1.1	-0.0	0.9	-0.7	-0.7				
18	m17	0.0	0.5	-0.2	-0.6	0.8	0.7	-0.4	-0.6			
19	v3a1	0.0	-0.6	-0.5	-0.4	-0.7	-0.5	0.3	0.4	-0.4		
20	v3a3	0.0	1.5	0.7	-0.3	-0.5	-0.2	-0.3	-0.7	-0.5	-0.4	
21	v3a9	0.0	2.1	1.6	0.2	1.1	-0.6	-0.6	-0.6	0.3	-0.0	0.4
22	v3a17	0.0	-0.6	-0.6	-0.4	-0.6	0.3	3.8	0.2	-0.4	1.3	-0.7
23	v3b4	0.0	-0.6	0.8	0.6	-0.6	-0.7	-0.6	0.3	-0.5	-0.7	1.7
24	v3b12	0.0	-0.1	0.4	----	----	----	-0.7	-0.3	0.7	-0.6	----
25	v3b15	0.0	2.0	0.6	-0.4	-0.6	0.5	-0.4	0.1	-0.7	-0.7	-0.6
26	v3b16	0.0	-0.7	-0.6	-0.1	-0.5	-0.5	0.2	0.8	-0.5	-0.5	-0.5
27	v5a7	0.0	-0.5	1.0	0.9	-0.7	1.8	0.4	-0.7	5.0	2.0	0.2
28	v5a10	0.0	-0.2	0.3	0.9	-0.5	0.1	3.5	-0.2	-0.6	-0.7	-0.6
29	v5a11	0.0	2.3	-0.6	0.6	1.2	-0.4	1.2	0.3	0.0	-0.5	0.9
30	v5a14	0.0	-0.7	0.6	-0.7	-0.6	-0.6	5.0	-0.7	4.1	-0.7	-0.1
31	v5a18	0.0	-0.6	-0.5	-0.7	2.5	-0.5	0.2	-0.7	2.0	0.4	-0.4
32	v5b2	0.0	1.4	-0.0	-0.6	-0.3	-0.6	-0.5	0.5	-0.7	3.2	0.4
33	v5b5	0.0	2.1	-0.6	-0.5	0.4	-0.7	-0.2	0.5	1.2	-0.6	-0.3
34	v5b6	0.0	-0.5	-0.6	-0.6	-0.3	0.9	1.6	0.1	1.3	-0.1	0.4
35	v5b8	0.0	-0.6	-0.0	-0.6	-0.4	-0.6	-0.7	0.3	-0.4	-0.5	-0.4
36	v5b13	0.0	-0.1	-0.6	1.1	0.7	-0.1	-0.5	-0.7	0.9	-0.3	0.2



Item	Label	Marginal χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.0										
22	v3a17	0.0	1.5									
23	v3b4	0.0	0.3	-0.7								
24	v3b12	0.0	----	-0.6	0.7							
25	v3b15	0.0	0.4	-0.6	0.5	-0.7						
26	v3b16	0.0	0.2	0.2	-0.3	----	-0.4					
27	v5a7	0.0	-0.5	-0.5	0.5	----	-0.7	-0.3				
28	v5a10	0.0	-0.4	2.9	-0.2	----	0.7	-0.7	-0.1			
29	v5a11	0.0	-0.7	0.2	0.3	----	0.9	0.2	-0.6	-0.7		
30	v5a14	0.0	0.9	1.2	-0.4	-0.2	-0.4	0.1	-0.1	0.0	-0.5	
31	v5a18	0.0	3.1	1.7	1.0	----	-0.0	-0.6	-0.0	-0.5	2.0	-0.7
32	v5b2	0.0	-0.7	-0.7	0.1	-0.7	-0.1	0.2	-0.6	-0.6	-0.5	0.2
33	v5b5	0.0	-0.2	-0.7	0.8	----	2.2	-0.1	-0.0	0.6	-0.4	-0.7
34	v5b6	0.0	-0.1	0.5	-0.7	0.6	-0.5	-0.6	-0.3	0.0	1.5	-0.7
35	v5b8	0.0	-0.5	-0.7	-0.7	1.8	4.3	0.3	-0.5	3.1	-0.6	-0.3
36	v5b13	0.0	-0.1	-0.7	-0.4	0.1	-0.6	-0.4	0.3	-0.1	-0.5	-0.6



Item	Label	Marginal χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	0.8				
33	v5b5	0.0	-0.5	-0.6			
34	v5b6	0.0	-0.4	-0.5	0.0		
35	v5b8	0.0	1.3	-0.2	-0.6	-0.5	
36	v5b13	0.0	2.1	-0.7	2.0	3.2	-0.7

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)

-2loglikelihood:	1511.04
Akaike Information Criterion (AIC):	1727.04
Bayesian Information Criterion (BIC):	1895.02

Summary of the Data and Control Parameters

Sample Size	35
Number of Items	36
Number of Dimensions	1

Appendix Y – IRTPRO output for Dimensionality & Model Selection (full instrument)

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - All Items; EFA(1)

Item	*Label*		*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e.*
1	m12	2	-0.76	0.23	1	0.1	0.18	0.14	0.24
2	m3	4	1.44	0.32	3	0.24	0.21	-0.17	0.14
3	m1	6	0.81	0.24	5	-0.05	0.18	0.06	0.22
4	m13	8	0.67	0.23	7	0.32	0.18	-0.48	0.29
5	m18	10	-0.1	0.18	9	0.06	0.17	0.57	2
6	m9	12	0.77	0.23	11	-0.23	0.18	0.29	0.24
7	m4	14	2.17	0.54	13	1.82	0.39	-0.84	0.13
8	m16	16	1.36	0.32	15	0.8	0.22	-0.59	0.17
9	m14	18	0.45	0.29	17	1.86	0.25	-4.15	2.57
10	m6	20	0.49	0.2	19	0.06	0.17	-0.12	0.35
11	m11	22	0.1	0.22	21	1.24	0.2	-12.32	27.31
12	m7	24	1.69	0.46	23	1.96	0.37	-1.16	0.21
13	m5	26	0.63	0.21	25	-0.3	0.18	0.47	0.31
14	m2	28	-0.02	0.18	27	-0.12	0.17	-6.49	62.78
15	m10	30	0.14	0.19	29	-0.23	0.17	1.69	2.53
16	m15	32	0.41	0.2	31	0.11	0.17	-0.27	0.43
17	m8	34	0.99	0.31	33	1.42	0.25	-1.44	0.38
18	m17	36	0.86	0.25	35	0.6	0.19	-0.7	0.26
19	v3a1	38	0.69	0.24	37	0.71	0.2	-1.03	0.39
20	v3a3	40	0.87	0.24	39	-0.31	0.19	0.36	0.24
21	v3a9	42	0.35	0.2	41	-0.4	0.18	1.17	0.84
22	v3a17	44	1.38	0.37	43	1.48	0.29	-1.07	0.23
23	v3b4	46	1.43	0.37	45	1.08	0.26	-0.76	0.18
24	v3b12	48	0.05	0.25	47	1.5	0.23	-28.49	136.2
25	v3b15	50	-0.47	0.22	49	-0.52	0.19	-1.12	0.62
26	v3b16	52	0.95	0.26	51	-0.1	0.19	0.1	0.21
27	v5a7	54	1.45	0.34	53	0.33	0.22	-0.23	0.15
28	v5a10	56	0.66	0.23	55	0.11	0.19	-0.17	0.28
29	v5a11	58	0.52	0.22	57	-0.1	0.18	0.19	0.37
30	v5a14	60	0.41	0.27	59	1.37	0.23	-3.36	2.12
31	v5a18	62	0.36	0.21	61	0.12	0.18	-0.34	0.52
32	v5b2	64	0.53	0.22	63	0.29	0.19	-0.54	0.39
33	v5b5	66	0.57	0.24	65	0.65	0.2	-1.14	0.53
34	v5b6	68	0.94	0.27	67	0.47	0.21	-0.5	0.23
35	v5b8	70	1.07	0.33	69	1.23	0.26	-1.16	0.31
36	v5b13	72	2.04	0.57	71	2.32	0.49	-1.14	0.18

Factor Loadings for Group 1

Item	Label	*λ1*	*s.e.*
1	m12	-0.41	0.18
2	m3	0.65	0.14
3	m1	0.43	0.17
4	m13	0.37	0.18
5	m18	-0.06	0.18
6	m9	0.41	0.17
7	m4	0.79	0.13
8	m16	0.62	0.15
9	m14	0.25	0.26
10	m6	0.28	0.18
11	m11	0.06	0.22
12	m7	0.71	0.16
13	m5	0.35	0.17
14	m2	-0.01	0.18
15	m10	0.08	0.18
16	m15	0.23	0.18
17	m8	0.5	0.2
18	m17	0.45	0.17
19	v3a1	0.38	0.19
20	v3a3	0.45	0.17
21	v3a9	0.2	0.19
22	v3a17	0.63	0.17
23	v3b4	0.64	0.16
24	v3b12	0.03	0.25
25	v3b15	-0.27	0.2
26	v3b16	0.49	0.17
27	v5a7	0.65	0.15
28	v5a10	0.36	0.19
29	v5a11	0.29	0.19
30	v5a14	0.23	0.25
31	v5a18	0.21	0.19
32	v5b2	0.3	0.19
33	v5b5	0.32	0.21
34	v5b6	0.48	0.18
35	v5b8	0.53	0.2
36	v5b13	0.77	0.15

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	5953.90 ± 0.75
Akaike Information Criterion (AIC):	6097.90 ± 0.75
Bayesian Information Criterion (BIC):	6313.21 ± 0.75

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	36
Number of Dimensions	1

IRTPRO Version 4.2

Output generated by IRTPRO estimation engine Version 5.20 (64-bit)

Project: Combined Sample - All Items; EFA(2)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*c*	*s.e.*
1	m12	2	-0.75	0.23		0	-----	1	0.1	0.18
2	m3	4	1.15	0.31	5	1.27	0.38	3	0.26	0.22
3	m1	7	0.47	0.24	8	1.21	0.34	6	-0.06	0.2
4	m13	10	0.6	0.22	11	0.27	0.23	9	0.32	0.18
5	m18	13	-0.29	0.2	14	0.32	0.22	12	0.05	0.17
6	m9	16	0.56	0.22	17	0.63	0.25	15	-0.23	0.18
7	m4	19	2.21	0.57	20	0.59	0.35	18	1.93	0.43
8	m16	22	1.51	0.35	23	0.18	0.27	21	0.87	0.24
9	m14	25	4.1	2.72	26	-4.47	2.71	24	6.77	3.92
10	m6	28	0.56	0.21	29	-0.01	0.22	27	0.06	0.17
11	m11	31	-0.08	0.23	32	0.35	0.26	30	1.27	0.21
12	m7	34	1.46	0.41	35	0.78	0.35	33	1.95	0.36
13	m5	37	0.42	0.21	38	0.63	0.26	36	-0.3	0.18
14	m2	40	-0.28	0.2	41	0.46	0.23	39	-0.13	0.17
15	m10	43	-0.15	0.2	44	0.6	0.24	42	-0.26	0.18
16	m15	46	0.23	0.2	47	0.47	0.23	45	0.11	0.17
17	m8	49	1.36	0.38	50	-0.23	0.29	48	1.63	0.31
18	m17	52	0.83	0.24	53	0.25	0.24	51	0.61	0.19
19	v3a1	55	0.54	0.23	56	0.39	0.25	54	0.71	0.2
20	v3a3	58	0.59	0.23	59	0.95	0.32	57	-0.33	0.2
21	v3a9	61	0.07	0.22	62	0.7	0.28	60	-0.43	0.19
22	v3a17	64	1.49	0.38	65	0.24	0.31	63	1.57	0.31
23	v3b4	67	1.33	0.35	68	0.48	0.3	66	1.09	0.27
24	v3b12	70	0.13	0.26	71	-0.13	0.29	69	1.51	0.23
25	v3b15	73	-0.36	0.22	74	-0.37	0.24	72	-0.53	0.19
26	v3b16	76	0.87	0.25	77	0.44	0.26	75	-0.09	0.2
27	v5a7	79	1.23	0.32	80	0.85	0.32	78	0.35	0.23
28	v5a10	82	0.51	0.22	83	0.45	0.25	81	0.12	0.19
29	v5a11	85	0.45	0.21	86	0.27	0.23	84	-0.1	0.18
30	v5a14	88	0.6	0.27	89	-0.27	0.28	87	1.44	0.25
31	v5a18	91	0.33	0.21	92	0.15	0.23	90	0.13	0.18
32	v5b2	94	0.32	0.22	95	0.52	0.26	93	0.29	0.19
33	v5b5	97	0.37	0.24	98	0.56	0.27	96	0.67	0.21
34	v5b6	100	0.8	0.27	101	0.5	0.27	99	0.47	0.21
35	v5b8	103	1.17	0.34	104	0.13	0.29	102	1.29	0.28
36	v5b13	106	1.68	0.54	107	1.36	0.47	105	2.39	0.52

Factor Loadings for Group 1

Item	Label	λ*1*	*s.e.*	λ2	s.e.
1	m12	-0.22	0.1	0.31	0.14
2	m3	-0.19	0.19	-0.71	0.14
3	m1	-0.37	0.2	-0.54	0.17
4	m13	0.05	0.21	-0.35	0.18
5	m18	-0.25	0.22	0	0.19
6	m9	-0.13	0.21	-0.45	0.18
7	m4	0.24	0.19	-0.73	0.14
8	m16	0.29	0.2	-0.56	0.16
9	m14	0.96	0.08	-0.02	0.22
10	m6	0.17	0.21	-0.23	0.18
11	m11	-0.2	0.26	-0.1	0.22
12	m7	0.05	0.22	-0.69	0.17
13	m5	-0.17	0.22	-0.4	0.18
14	m2	-0.3	0.21	-0.05	0.19
15	m10	-0.33	0.21	-0.16	0.19
16	m15	-0.15	0.21	-0.27	0.18
17	m8	0.42	0.21	-0.4	0.2
18	m17	0.12	0.21	-0.42	0.18
19	v3a1	-0.03	0.23	-0.37	0.2
20	v3a3	-0.25	0.22	-0.53	0.18
21	v3a9	-0.31	0.23	-0.28	0.19
22	v3a17	0.26	0.22	-0.57	0.18
23	v3b4	0.14	0.22	-0.6	0.18
24	v3b12	0.11	0.29	-0.01	0.26
25	v3b15	0.07	0.23	0.29	0.2
26	v3b16	0.04	0.22	-0.49	0.18
27	v5a7	-0.03	0.21	-0.66	0.16
28	v5a10	-0.06	0.22	-0.38	0.19
29	v5a11	0	0.22	-0.29	0.2
30	v5a14	0.31	0.25	-0.15	0.25
31	v5a18	0.03	0.23	-0.2	0.2
32	v5b2	-0.16	0.23	-0.33	0.2
33	v5b5	-0.16	0.23	-0.36	0.21
34	v5b6	0	0.22	-0.49	0.19
35	v5b8	0.25	0.24	-0.47	0.21
36	v5b13	-0.1	0.22	-0.8	0.15

Factor Correlation Matrix

	*λ*1	*λ*2
*λ*1	1	-0.16
*λ*2	-0.16	1

Likelihood-based Values and Goodness of Fit Statistics	
Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	5881.59 ± 1.31
Akaike Information Criterion (AIC):	6095.59 ± 1.31
Bayesian Information Criterion (BIC):	6415.56 ± 1.31

Summary of the Data and Control Parameters	
Sample Size	147
Number of Items	36
Number of Dimensions	2

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - All Items; EFA(3)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*a3*	*s.e.*		*c*	*s.e.*
1	m12	2	-1.04	0.29		0	-----		0	-----	1	0.11	0.19
2	m3	4	1.39	0.38	5	1.28	0.36		0	-----	3	0.25	0.22
3	m1	7	0.46	0.25	8	1.26	0.35	9	0.01	0.25	6	-0.07	0.2
4	m13	11	0.57	0.23	12	0.32	0.22	13	0.15	0.22	10	0.31	0.18
5	m18	15	-0.5	0.24	16	0.46	0.26	17	0.12	0.24	14	0.06	0.18
6	m9	19	0.3	0.22	20	0.89	0.29	21	0.51	0.28	18	-0.25	0.19
7	m4	23	2.2	0.63	24	0.65	0.37	25	0.9	0.37	22	1.97	0.45
8	m16	27	1.34	0.36	28	0.26	0.27	29	0.73	0.29	26	0.85	0.24
9	m14	31	2.64	2.01	32	-3.59	2.61	33	3.64	2.87	30	6.3	4.33
10	m6	35	0.72	0.25	36	-0.12	0.22	37	0.03	0.24	34	0.06	0.18
11	m11	39	0.26	0.26	40	0.21	0.26	41	-0.62	0.3	38	1.36	0.24
12	m7	43	1.82	0.56	44	0.7	0.37	45	0.2	0.33	42	2.1	0.43
13	m5	47	0.42	0.22	48	0.59	0.25	49	0.06	0.22	46	-0.31	0.18
14	m2	51	-0.17	0.21	52	0.39	0.22	53	-0.31	0.23	50	-0.13	0.17
15	m10	55	-0.43	0.26	56	0.88	0.3	57	0.19	0.25	54	-0.28	0.19
16	m15	59	0.12	0.21	60	0.53	0.25	61	0.15	0.22	58	0.11	0.17
17	m8	63	0.81	0.33	64	0.14	0.32	65	1.56	0.5	62	1.8	0.37
18	m17	67	0.83	0.27	68	0.23	0.24	69	0.24	0.23	66	0.6	0.19
19	v3a1	71	0.5	0.25	72	0.43	0.26	73	0.17	0.23	70	0.7	0.2
20	v3a3	75	1.03	0.32	76	0.83	0.32	77	-0.44	0.27	74	-0.36	0.21
21	v3a9	79	0.02	0.23	80	0.74	0.28	81	-0.02	0.23	78	-0.43	0.19
22	v3a17	83	0.92	0.36	84	0.84	0.38	85	1.67	0.48	82	1.84	0.41
23	v3b4	87	1.49	0.41	88	0.4	0.29	89	0.42	0.29	86	1.13	0.28
24	v3b12	91	-0.04	0.27	92	0.01	0.28	93	0.25	0.3	90	1.51	0.23
25	v3b15	95	-0.34	0.23	96	-0.34	0.25	97	-0.1	0.23	94	-0.52	0.19
26	v3b16	99	0.91	0.28	100	0.4	0.26	101	0.21	0.24	98	-0.1	0.2
27	v5a7	103	1.12	0.32	104	0.94	0.33	105	0.44	0.26	102	0.32	0.23
28	v5a10	107	0.1	0.25	108	0.89	0.32	109	0.76	0.31	106	0.11	0.21
29	v5a11	111	0.61	0.24	112	0.23	0.24	113	-0.12	0.23	110	-0.09	0.19
30	v5a14	115	0.4	0.28	116	-0.18	0.29	117	0.51	0.29	114	1.42	0.25
31	v5a18	119	0.08	0.21	120	0.33	0.24	121	0.43	0.24	118	0.11	0.19
32	v5b2	123	0.46	0.25	124	0.45	0.25	125	-0.16	0.24	122	0.3	0.2
33	v5b5	127	0.41	0.25	128	0.55	0.26	129	-0.04	0.25	126	0.66	0.21
34	v5b6	131	0.86	0.28	132	0.5	0.27	133	0.1	0.25	130	0.47	0.21
35	v5b8	135	0.52	0.54	136	1.71	1.01	137	3.83	2.14	134	2.78	1.35
36	v5b13	139	1.49	0.54	140	1.52	0.53	141	0.56	0.39	138	2.36	0.52

Item	Label	*λ1*	*s.e.*	λ2	s.e.	λ3	s.e.
1	m12	-0.19	0.06	-0.51	0.18	0.21	0.07
2	m3	-0.04	0.11	0.72	0.15	0.2	0.17
3	m1	-0.2	0.22	0.41	0.2	0.4	0.2
4	m13	-0.05	0.23	0.34	0.22	-0.01	0.21
5	m18	-0.27	0.25	-0.2	0.22	0.29	0.23
6	m9	-0.39	0.25	0.23	0.21	0.2	0.21
7	m4	-0.12	0.22	0.71	0.18	-0.24	0.2
8	m16	-0.14	0.23	0.51	0.2	-0.27	0.2
9	m14	-0.13	0.25	0.03	0.26	-0.93	0.1
10	m6	0.16	0.25	0.36	0.22	-0.22	0.21
11	m11	0.31	0.27	0.29	0.26	0.17	0.25
12	m7	0.06	0.27	0.76	0.2	-0.09	0.22
13	m5	-0.1	0.24	0.33	0.21	0.16	0.21
14	m2	0.03	0.23	0.04	0.21	0.29	0.21
15	m10	-0.38	0.25	-0.09	0.23	0.42	0.21
16	m15	-0.2	0.23	0.14	0.21	0.18	0.22
17	m8	-0.5	0.25	0.14	0.24	-0.35	0.21
18	m17	-0.02	0.24	0.43	0.22	-0.13	0.21
19	v3a1	-0.1	0.25	0.32	0.23	0.05	0.24
20	v3a3	0.18	0.23	0.66	0.19	0.21	0.21
21	v3a9	-0.17	0.24	0.16	0.23	0.33	0.23
22	v3a17	-0.6	0.2	0.25	0.22	-0.13	0.21
23	v3b4	-0.02	0.24	0.63	0.2	-0.19	0.21
24	v3b12	-0.15	0.33	-0.06	0.3	-0.05	0.29
25	v3b15	0.07	0.25	-0.24	0.23	-0.06	0.24
26	v3b16	-0.03	0.24	0.49	0.21	-0.06	0.22
27	v5a7	-0.2	0.22	0.57	0.19	0.07	0.2
28	v5a10	-0.53	0.24	0.09	0.23	0.19	0.21
29	v5a11	0.12	0.25	0.4	0.22	0	0.22
30	v5a14	-0.13	0.3	0.09	0.28	-0.29	0.27
31	v5a18	-0.3	0.24	0.03	0.23	0.04	0.23
32	v5b2	0.06	0.25	0.36	0.23	0.14	0.23
33	v5b5	-0.04	0.25	0.33	0.24	0.17	0.23
34	v5b6	-0.01	0.24	0.5	0.22	0.01	0.22
35	v5b8	-0.92	0.17	-0.02	0.22	-0.07	0.19
36	v5b13	-0.25	0.27	0.66	0.22	0.16	0.21

Factor Correlation Matrix

	*λ*1	*λ*2	*λ3*
*λ*1	1	-0.41	0.16
*λ*2	-0.41	1	-0.09
λ3	0.16	-0.09	1

Likelihood-based Values and Goodness of Fit Statistics	
Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	5819.86 ± 1.84
Akaike Information Criterion (AIC):	6101.86 ± 1.84
Bayesian Information Criterion (BIC):	6523.51 ± 1.84

Summary of the Data and Control Parameters	
Sample Size	147
Number of Items	36
Number of Dimensions	3

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - All Items; CFA(2)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*c*	*s.e.*
1	m12	2	-0.71	0.23		0	-----	1	0.1	0.18
2	m3	4	1.65	0.38		0	-----	3	0.27	0.22
3	m1	6	0.94	0.25		0	-----	5	-0.05	0.18
4	m13	8	0.69	0.23		0	-----	7	0.32	0.18
5	m18		0	-----	10	0.36	0.25	9	0.06	0.17
6	m9	12	0.76	0.23		0	-----	11	-0.23	0.18
7	m4	14	2.11	0.54		0	-----	13	1.81	0.38
8	m16		0	-----	16	1.37	0.5	15	0.79	0.25
9	m14	18	0.32	0.29		0	-----	17	1.83	0.25
10	m6	20	0.39	0.2		0	-----	19	0.06	0.17
11	m11	22	0.23	0.23		0	-----	21	1.25	0.2
12	m7	24	1.98	0.54		0	-----	23	2.16	0.44
13	m5	26	0.61	0.22		0	-----	25	-0.29	0.18
14	m2	28	0.02	0.19		0	-----	27	-0.12	0.17
15	m10	30	0.2	0.19		0	-----	29	-0.23	0.17
16	m15		0	-----	32	0.61	0.28	31	0.12	0.18
17	m8	34	0.81	0.28		0	-----	33	1.36	0.23
18	m17		0	-----	36	0.69	0.29	35	0.57	0.19
19	v3a1	38	0.66	0.24		0	-----	37	0.71	0.2
20	v3a3	40	0.81	0.24		0	-----	39	-0.3	0.19
21	v3a9	42	0.4	0.21		0	-----	41	-0.4	0.18
22	v3a17		0	-----	44	1.15	0.44	43	1.33	0.28
23	v3b4	46	1.24	0.34		0	-----	45	1.02	0.25
24	v3b12	48	0.08	0.26		0	-----	47	1.5	0.23
25	v3b15		0	-----	50	-0.36	0.25	49	-0.5	0.19
26	v3b16		0	-----	52	1.48	0.52	51	-0.16	0.23
27	v5a7	54	1.43	0.35		0	-----	53	0.34	0.23
28	v5a10	56	0.65	0.24		0	-----	55	0.12	0.19
29	v5a11	58	0.49	0.22		0	-----	57	-0.09	0.18
30	v5a14	60	0.39	0.26		0	-----	59	1.37	0.23
31	v5a18		0	-----	62	0.49	0.25	61	0.11	0.18
32	v5b2	64	0.57	0.23		0	-----	63	0.3	0.19
33	v5b5	66	0.59	0.24		0	-----	65	0.66	0.2
34	v5b6	68	0.86	0.27		0	-----	67	0.46	0.21
35	v5b8	70	0.92	0.32		0	-----	69	1.19	0.25
36	v5b13	72	2.18	0.7		0	-----	71	2.47	0.59

Factor Loadings for Group 1

Item	Label	λ*1*	*s.e.*	λ2	s.e.
1	m12	-0.39	0.18	0	0
2	m3	0.7	0.14	0	0
3	m1	0.48	0.16	0	0
4	m13	0.38	0.18	0	0
5	m18	0	0	0.21	0.23
6	m9	0.41	0.17	0	0
7	m4	0.78	0.13	0	0
8	m16	0	0	0.63	0.24
9	m14	0.19	0.27	0	0
10	m6	0.22	0.18	0	0
11	m11	0.13	0.23	0	0
12	m7	0.76	0.15	0	0
13	m5	0.34	0.18	0	0
14	m2	0.01	0.19	0	0
15	m10	0.12	0.19	0	0
16	m15	0	0	0.34	0.23
17	m8	0.43	0.2	0	0
18	m17	0	0	0.38	0.23
19	v3a1	0.36	0.2	0	0
20	v3a3	0.43	0.18	0	0
21	v3a9	0.23	0.2	0	0
22	v3a17	0	0	0.56	0.25
23	v3b4	0.59	0.18	0	0
24	v3b12	0.05	0.26	0	0
25	v3b15	0	0	-0.21	0.23
26	v3b16	0	0	0.66	0.22
27	v5a7	0.64	0.16	0	0
28	v5a10	0.35	0.19	0	0
29	v5a11	0.28	0.19	0	0
30	v5a14	0.22	0.24	0	0
31	v5a18	0	0	0.28	0.22
32	v5b2	0.32	0.2	0	0
33	v5b5	0.33	0.2	0	0
34	v5b6	0.45	0.19	0	0
35	v5b8	0.48	0.21	0	0
36	v5b13	0.79	0.16	0	0

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	6007.63 ± 1.10
Akaike Information Criterion (AIC):	61051.63 ± 1.10
Bayesian Information Criterion (BIC):	6366.94 ± 1.10

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	36
Number of Dimensions	2

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - All Items; CFA(2) by item type

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*c*	*s.e.*
1	m12	2	-0.66	0.23		0	-----	1	0.11	0.18
2	m3	4	1.54	0.37		0	-----	3	0.23	0.21
3	m1	6	0.8	0.24		0	-----	5	-0.06	0.18
4	m13	8	0.57	0.22		0	-----	7	0.31	0.18
5	m18	10	-0.22	0.2		0	-----	9	0.06	0.17
6	m9	12	0.68	0.23		0	-----	11	-0.23	0.18
7	m4	14	2.23	0.68		0	-----	13	1.86	0.45
8	m16	16	1.19	0.32		0	-----	15	0.75	0.22
9	m14	18	0.62	0.32		0	-----	17	1.92	0.27
10	m6	20	0.59	0.22		0	-----	19	0.06	0.17
11	m11	22	0.29	0.24		0	-----	21	1.26	0.2
12	m7	24	1.99	0.62		0	-----	23	2.17	0.48
13	m5	26	0.64	0.23		0	-----	25	-0.3	0.18
14	m2	28	-0.13	0.19		0	-----	27	-0.12	0.17
15	m10	30	0.01	0.2		0	-----	29	-0.23	0.17
16	m15	32	0.15	0.2		0	-----	31	0.11	0.17
17	m8	34	0.93	0.31		0	-----	33	1.4	0.25
18	m17	36	0.76	0.27		0	-----	35	0.58	0.19
19	v3a1		0	-----	38	0.66	0.27	37	0.68	0.2
20	v3a3		0	-----	40	0.86	0.29	39	-0.36	0.19
21	v3a9		0	-----	42	0.43	0.23	41	-0.42	0.18
22	v3a17		0	-----	44	1.14	0.37	43	1.32	0.26
23	v3b4		0	-----	46	1.09	0.34	45	0.91	0.23
24	v3b12		0	-----	48	0.16	0.28	47	1.5	0.23
25	v3b15		0	-----	50	-0.55	0.24	49	-0.51	0.19
26	v3b16		0	-----	52	1.16	0.32	51	-0.15	0.21
27	v5a7		0	-----	54	1.43	0.38	53	0.21	0.22
28	v5a10		0	-----	56	0.66	0.26	55	0.07	0.19
29	v5a11		0	-----	58	0.67	0.26	57	-0.14	0.19
30	v5a14		0	-----	60	0.39	0.27	59	1.34	0.23
31	v5a18		0	-----	62	0.48	0.23	61	0.1	0.18
32	v5b2		0	-----	64	0.46	0.24	63	0.25	0.19
33	v5b5		0	-----	66	0.56	0.26	65	0.61	0.2
34	v5b6		0	-----	68	1.13	0.34	67	0.42	0.22
35	v5b8		0	-----	70	0.94	0.32	69	1.12	0.24
36	v5b13		0	-----	72	1.86	0.61	71	2.07	0.47

Factor Loadings for Group 1

Item	Label	λ*1*	*s.e.*	λ2	s.e.
1	m12	-0.36	0.18	0	0
2	m3	0.67	0.15	0	0
3	m1	0.42	0.18	0	0
4	m13	0.32	0.19	0	0
5	m18	-0.13	0.19	0	0
6	m9	0.37	0.18	0	0
7	m4	0.8	0.15	0	0
8	m16	0.57	0.18	0	0
9	m14	0.34	0.26	0	0
10	m6	0.33	0.19	0	0
11	m11	0.17	0.23	0	0
12	m7	0.76	0.17	0	0
13	m5	0.35	0.19	0	0
14	m2	-0.08	0.19	0	0
15	m10	0.01	0.2	0	0
16	m15	0.09	0.2	0	0
17	m8	0.48	0.21	0	0
18	m17	0.41	0.2	0	0
19	v3a1	0	0	0.36	0.21
20	v3a3	0	0	0.45	0.2
21	v3a9	0	0	0.24	0.21
22	v3a17	0	0	0.56	0.21
23	v3b4	0	0	0.54	0.2
24	v3b12	0	0	0.09	0.28
25	v3b15	0	0	-0.31	0.21
26	v3b16	0	0	0.56	0.18
27	v5a7	0	0	0.64	0.17
28	v5a10	0	0	0.36	0.21
29	v5a11	0	0	0.37	0.21
30	v5a14	0	0	0.22	0.25
31	v5a18	0	0	0.27	0.21
32	v5b2	0	0	0.26	0.21
33	v5b5	0	0	0.32	0.22
34	v5b6	0	0	0.55	0.2
35	v5b8	0	0	0.48	0.22
36	v5b13	0	0	0.74	0.19

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	6032.4 ± 1.16
Akaike Information Criterion (AIC):	6176.40 ± 1.16
Bayesian Information Criterion (BIC):	6391.71 ± 1.16

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	36
Number of Dimensions	2

IRTPRO Version 4.2

Output generated by IRTPRO estimation engine Version 5.20 (64-bit)

Project: Combined Sample - All Items; Bifactor

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*a3*	*s.e.*		*c*	*s.e.*
1	m12	2	-0.77	0.22	3	-0.16	0.23		0	-----	1	0.1	0.18
2	m3	5	1.38	0.34	6	1.15	0.35		0	-----	4	0.26	0.22
3	m1	8	0.68	0.27	9	1.26	0.37		0	-----	7	-0.07	0.21
4	m13	11	0.67	0.22	12	0.18	0.22		0	-----	10	0.32	0.18
5	m18	14	-0.97	1.01		0	-----	15	4.28	4.2	13	0.18	0.55
6	m9	17	0.66	0.22	18	0.48	0.25		0	-----	16	-0.23	0.18
7	m4	20	2.32	0.55	21	0.28	0.34		0	-----	19	1.93	0.41
8	m16	23	1.48	0.34		0	-----	24	0.11	0.28	22	0.85	0.23
9	m14	26	4.13	7.97	27	-5.94	10.69		0	-----	25	7.93	13.55
10	m6	29	0.54	0.21	30	-0.05	0.21		0	-----	28	0.06	0.17
11	m11	32	-0.06	0.24	33	0.59	0.28		0	-----	31	1.33	0.22
12	m7	35	1.66	0.46	36	0.64	0.35		0	-----	34	2	0.37
13	m5	38	0.5	0.21	39	0.59	0.26		0	-----	37	-0.31	0.18
14	m2	41	-0.19	0.2	42	0.43	0.23		0	-----	40	-0.13	0.17
15	m10	44	-0.04	0.2	45	0.56	0.24		0	-----	43	-0.25	0.18
16	m15	47	0.32	0.2		0	-----	48	0.25	0.22	46	0.11	0.17
17	m8	50	1.35	0.38	51	-0.45	0.3		0	-----	49	1.64	0.3
18	m17	53	0.86	0.25		0	-----	54	0.11	0.23	52	0.6	0.19
19	v3a1	56	0.63	0.24	57	0.29	0.25		0	-----	55	0.71	0.2
20	v3a3	59	0.75	0.25	60	0.84	0.3		0	-----	58	-0.33	0.2
21	v3a9	62	0.2	0.21	63	0.64	0.27		0	-----	61	-0.42	0.19
22	v3a17	65	1.6	0.44		0	-----	66	0.6	0.33	64	1.65	0.34
23	v3b4	68	1.46	0.36	69	0.22	0.29		0	-----	67	1.1	0.27
24	v3b12	71	0.08	0.26	72	-0.08	0.29		0	-----	70	1.5	0.23
25	v3b15	74	-0.44	0.23		0	-----	75	-0.22	0.23	73	-0.52	0.19
26	v3b16	77	0.93	0.26		0	-----	78	0.32	0.26	76	-0.11	0.2
27	v5a7	80	1.37	0.34	81	0.58	0.29		0	-----	79	0.34	0.23
28	v5a10	83	0.6	0.23	84	0.37	0.24		0	-----	82	0.12	0.19
29	v5a11	86	0.48	0.21	87	0.28	0.24		0	-----	85	-0.1	0.19
30	v5a14	89	0.58	0.28	90	-0.44	0.29		0	-----	88	1.45	0.25
31	v5a18	92	0.49	0.31		0	-----	93	1.9	0.72	91	0.12	0.26
32	v5b2	95	0.41	0.23	96	0.52	0.26		0	-----	94	0.3	0.2
33	v5b5	98	0.44	0.24	99	0.55	0.27		0	-----	97	0.68	0.21
34	v5b6	101	0.88	0.27	102	0.31	0.26		0	-----	100	0.47	0.21
35	v5b8	104	1.17	0.34	105	-0.09	0.28		0	-----	103	1.28	0.27
36	v5b13	107	1.85	0.56	108	0.93	0.43		0	-----	106	2.32	0.51

Factor Loadings for Group 1

Item	Label	λ1	*s.e.*	λ2	s.e.	λ3	s.e.
1	m12	-0.41	0.17	-0.09	0.21	0	0
2	m3	0.56	0.14	0.46	0.17	0	0
3	m1	0.31	0.18	0.57	0.18	0	0
4	m13	0.36	0.18	0.1	0.2	0	0
5	m18	-0.21	0.19	0	0	0.91	0.21
6	m9	0.35	0.17	0.25	0.21	0	0
7	m4	0.8	0.11	0.1	0.19	0	0
8	m16	0.66	0.15	0	0	0.05	0.21
9	m14	0.56	0.26	-0.8	0.16	0	0
10	m6	0.3	0.18	-0.03	0.2	0	0
11	m11	-0.03	0.22	0.33	0.24	0	0
12	m7	0.67	0.17	0.26	0.21	0	0
13	m5	0.27	0.18	0.31	0.21	0	0
14	m2	-0.11	0.19	0.24	0.21	0	0
15	m10	-0.02	0.19	0.31	0.21	0	0
16	m15	0.18	0.19	0	0	0.14	0.21
17	m8	0.61	0.18	-0.2	0.21	0	0
18	m17	0.45	0.18	0	0	0.06	0.21
19	v3a1	0.34	0.19	0.16	0.22	0	0
20	v3a3	0.37	0.17	0.41	0.2	0	0
21	v3a9	0.11	0.2	0.35	0.22	0	0
22	v3a17	0.66	0.17	0	0	0.25	0.21
23	v3b4	0.65	0.16	0.1	0.22	0	0
24	v3b12	0.05	0.26	-0.05	0.29	0	0
25	v3b15	-0.25	0.2	0	0	-0.12	0.22
26	v3b16	0.47	0.17	0	0	0.16	0.22
27	v5a7	0.61	0.15	0.26	0.2	0	0
28	v5a10	0.33	0.19	0.2	0.22	0	0
29	v5a11	0.27	0.19	0.16	0.22	0	0
30	v5a14	0.32	0.23	-0.24	0.25	0	0
31	v5a18	0.19	0.18	0	0	0.73	0.21
32	v5b2	0.23	0.2	0.28	0.22	0	0
33	v5b5	0.24	0.21	0.3	0.23	0	0
34	v5b6	0.45	0.18	0.16	0.22	0	0
35	v5b8	0.57	0.19	-0.04	0.23	0	0
36	v5b13	0.69	0.17	0.35	0.21	0	0

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)

-2loglikelihood:	5849.53 ± 1.92
Akaike Information Criterion (AIC):	6065.53 ± 1.92
Bayesian Information Criterion (BIC):	6388.50 ± 1.92

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	36
Number of Dimensions	3

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - All Items; Bifactor by item type

Item	***Label***		***a1***	***s.e.***		***a2***	***s.e.***		***a3***	***s.e.***		***c***	***s.e.***
1	m12	2	-0.74	0.22	3	0.34	0.25		0	-----	1	0.09	0.18
2	m3	5	1.56	0.38	6	0.31	0.3		0	-----	4	0.27	0.21
3	m1	8	1.15	0.35	9	1.06	0.46		0	-----	7	-0.05	0.22
4	m13	11	0.67	0.22	12	-0.17	0.23		0	-----	10	0.33	0.18
5	m18	14	-0.08	0.19	15	0.52	0.25		0	-----	13	0.05	0.17
6	m9	17	0.88	0.25	18	0.47	0.26		0	-----	16	-0.23	0.19
7	m4	20	2.2	0.56	21	-0.76	0.4		0	-----	19	1.96	0.43
8	m16	23	1.35	0.33	24	-0.64	0.32		0	-----	22	0.87	0.24
9	m14	26	0.61	0.49	27	-2.58	1.18		0	-----	25	3.39	1.11
10	m6	29	0.48	0.2	30	-0.39	0.25		0	-----	28	0.07	0.18
11	m11	32	0.11	0.22	33	0.28	0.28		0	-----	31	1.26	0.2
12	m7	35	1.68	0.46	36	-0.27	0.35		0	-----	34	1.99	0.37
13	m5	38	0.67	0.22	39	0.31	0.25		0	-----	37	-0.29	0.18
14	m2	41	0.01	0.19	42	0.59	0.28		0	-----	40	-0.13	0.18
15	m10	44	0.22	0.21	45	0.73	0.29		0	-----	43	-0.26	0.18
16	m15	47	0.44	0.21	48	0.36	0.25		0	-----	46	0.12	0.17
17	m8	50	0.99	0.31	51	-0.68	0.33		0	-----	49	1.54	0.28
18	m17	53	0.84	0.25	54	-0.28	0.26		0	-----	52	0.62	0.19
19	v3a1	56	1.24	0.47		0	-----	57	2.36	0.94	55	1.3	0.46
20	v3a3	59	0.92	0.26		0	-----	60	-0.34	0.27	58	-0.31	0.19
21	v3a9	62	0.38	0.2		0	-----	63	0.25	0.24	61	-0.41	0.18
22	v3a17	65	1.37	0.37		0	-----	66	-0.32	0.31	64	1.51	0.3
23	v3b4	68	1.42	0.36		0	-----	69	0.34	0.3	67	1.1	0.26
24	v3b12	71	0.05	0.26		0	-----	72	0.47	0.32	70	1.55	0.24
25	v3b15	74	-0.49	0.22		0	-----	75	0.1	0.25	73	-0.54	0.19
26	v3b16	77	0.95	0.26		0	-----	78	0.08	0.27	76	-0.08	0.19
27	v5a7	80	1.77	0.45		0	-----	81	-0.82	0.38	79	0.45	0.26
28	v5a10	83	0.68	0.23		0	-----	84	-0.13	0.26	82	0.13	0.19
29	v5a11	86	0.52	0.21		0	-----	87	0.15	0.24	85	-0.1	0.19
30	v5a14	89	0.39	0.27		0	-----	90	0.56	0.31	88	1.42	0.25
31	v5a18	92	0.39	0.21		0	-----	93	-0.26	0.25	91	0.15	0.18
32	v5b2	95	0.85	0.4		0	-----	96	2.1	0.97	94	0.36	0.3
33	v5b5	98	0.58	0.24		0	-----	99	-0.17	0.27	97	0.67	0.21
34	v5b6	101	0.96	0.27		0	-----	102	-0.2	0.27	100	0.49	0.21
35	v5b8	104	1.05	0.33		0	-----	105	-0.36	0.3	103	1.28	0.27
36	v5b13	107	2.18	0.64		0	-----	108	0.34	0.38	106	2.45	0.55

Factor Loadings for Group 1

Item	Label	λ*1*	*s.e.*	λ2	s.e.	λ3	s.e.
1	m12	-0.39	0.17	0.18	0.22	0	0
2	m3	0.67	0.15	0.13	0.21	0	0
3	m1	0.5	0.17	0.46	0.24	0	0
4	m13	0.36	0.18	-0.09	0.21	0	0
5	m18	-0.04	0.18	0.29	0.22	0	0
6	m9	0.45	0.17	0.24	0.21	0	0
7	m4	0.76	0.13	-0.26	0.21	0	0
8	m16	0.6	0.15	-0.28	0.22	0	0
9	m14	0.19	0.23	-0.82	0.2	0	0
10	m6	0.26	0.18	-0.21	0.22	0	0
11	m11	0.06	0.22	0.16	0.27	0	0
12	m7	0.7	0.17	-0.11	0.24	0	0
13	m5	0.36	0.18	0.17	0.22	0	0
14	m2	0	0.18	0.33	0.24	0	0
15	m10	0.12	0.18	0.39	0.22	0	0
16	m15	0.25	0.18	0.2	0.22	0	0
17	m8	0.48	0.19	-0.33	0.23	0	0
18	m17	0.44	0.18	-0.15	0.23	0	0
19	v3a1	0.39	0.17	0	0	0.75	0.18
20	v3a3	0.47	0.17	0	0	-0.18	0.22
21	v3a9	0.21	0.19	0	0	0.14	0.23
22	v3a17	0.62	0.17	0	0	-0.14	0.23
23	v3b4	0.63	0.16	0	0	0.15	0.22
24	v3b12	0.03	0.25	0	0	0.26	0.29
25	v3b15	-0.28	0.2	0	0	0.06	0.24
26	v3b16	0.49	0.17	0	0	0.04	0.23
27	v5a7	0.68	0.14	0	0	-0.32	0.2
28	v5a10	0.37	0.19	0	0	-0.07	0.24
29	v5a11	0.29	0.19	0	0	0.08	0.23
30	v5a14	0.21	0.24	0	0	0.3	0.26
31	v5a18	0.22	0.19	0	0	-0.15	0.23
32	v5b2	0.3	0.18	0	0	0.74	0.23
33	v5b5	0.32	0.21	0	0	-0.09	0.25
34	v5b6	0.49	0.18	0	0	-0.1	0.23
35	v5b8	0.52	0.2	0	0	-0.18	0.24
36	v5b13	0.78	0.15	0	0	0.12	0.23

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	5863.76 ± 1.74
Akaike Information Criterion (AIC):	6079.76 ± 1.74
Bayesian Information Criterion (BIC):	6402.73 ± 1.74

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	36
Number of Dimensions	3

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - All Items; Rasch

Item	*Label*	*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e.*
1	m12	1	-----	1	0.11	0.18	-0.11	0.18
2	m3	1	-----	2	0.17	0.18	-0.17	0.18
3	m1	1	-----	3	-0.05	0.18	0.05	0.18
4	m13	1	-----	4	0.32	0.18	-0.32	0.18
5	m18	1	-----	5	0.06	0.18	-0.06	0.18
6	m9	1	-----	6	-0.23	0.18	0.23	0.18
7	m4	1	-----	7	1.17	0.2	-1.17	0.2
8	m16	1	-----	8	0.65	0.19	-0.65	0.19
9	m14	1	-----	9	1.96	0.25	-1.96	0.25
10	m6	1	-----	10	0.06	0.18	-0.06	0.18
11	m11	1	-----	11	1.37	0.21	-1.37	0.21
12	m7	1	-----	12	1.49	0.22	-1.49	0.22
13	m5	1	-----	13	-0.3	0.19	0.3	0.19
14	m2	1	-----	14	-0.14	0.18	0.14	0.18
15	m10	1	-----	15	-0.26	0.18	0.26	0.18
16	m15	1	-----	16	0.12	0.18	-0.12	0.18
17	m8	1	-----	17	1.33	0.21	-1.33	0.21
18	m17	1	-----	18	0.58	0.19	-0.58	0.19
19	v3a1	1	-----	19	0.72	0.2	-0.72	0.2
20	v3a3	1	-----	20	-0.31	0.19	0.31	0.19
21	v3a9	1	-----	21	-0.42	0.19	0.42	0.19
22	v3a17	1	-----	22	1.2	0.21	-1.2	0.21
23	v3b4	1	-----	23	0.84	0.2	-0.84	0.2
24	v3b12	1	-----	24	1.65	0.24	-1.65	0.24
25	v3b15	1	-----	25	-0.54	0.2	0.54	0.2
26	v3b16	1	-----	26	-0.11	0.19	0.11	0.19
27	v5a7	1	-----	27	0.21	0.2	-0.21	0.2
28	v5a10	1	-----	28	0.1	0.2	-0.1	0.2
29	v5a11	1	-----	29	-0.11	0.2	0.11	0.2
30	v5a14	1	-----	30	1.46	0.23	-1.46	0.23
31	v5a18	1	-----	31	0.14	0.2	-0.14	0.2
32	v5b2	1	-----	32	0.29	0.2	-0.29	0.2
33	v5b5	1	-----	33	0.66	0.21	-0.66	0.21
34	v5b6	1	-----	34	0.4	0.2	-0.4	0.2
35	v5b8	1	-----	35	1.09	0.22	-1.09	0.22
36	v5b13	1	-----	36	1.52	0.24	-1.52	0.24

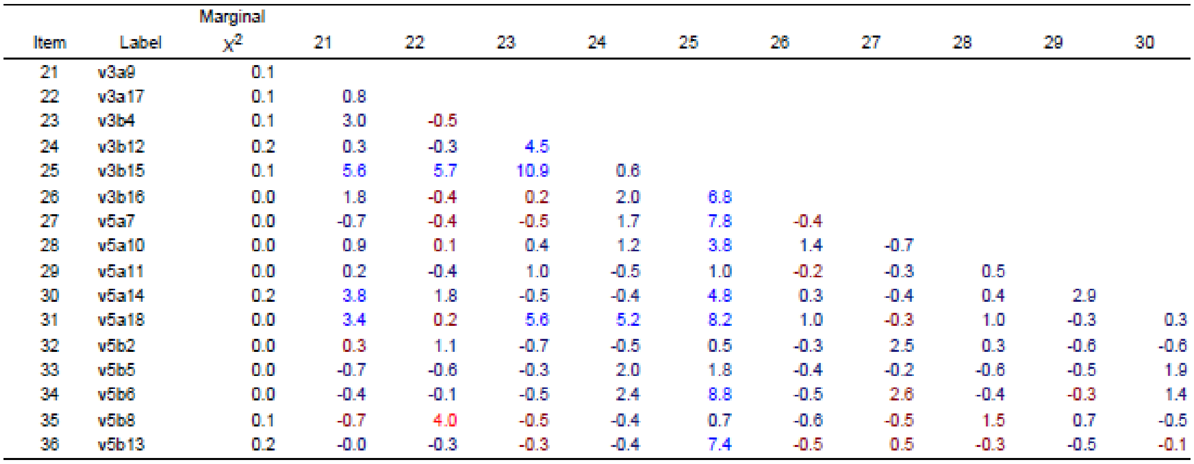
S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	Probability
1	m12	59.01	13	0.0001
2	m3	18.9	14	0.1683
3	m1	11.73	14	0.6293
4	m13	16.52	16	0.4188
5	m18	33.97	15	0.0034
6	m9	11.31	14	0.6626
7	m4	14.43	12	0.2732
8	m16	28.95	15	0.0163
9	m14	12.69	9	0.1767
10	m6	21.36	15	0.1254
11	m11	17.53	13	0.1756
12	m7	13.35	11	0.2704
13	m5	11.3	13	0.5867
14	m2	25.56	15	0.0428
15	m10	25.41	15	0.0445
16	m15	14.71	14	0.3997
17	m8	12.15	12	0.4357
18	m17	12.36	12	0.419
19	v3a1	11.02	13	0.6102
20	v3a3	17.09	15	0.3129
21	v3a9	16.71	13	0.2123
22	v3a17	11.85	11	0.3769
23	v3b4	10.12	12	0.6065
24	v3b12	21.67	11	0.0269
25	v3b15	49.09	16	0.0001
26	v3b16	12.03	15	0.6775
27	v5a7	17.38	15	0.2958
28	v5a10	23.94	14	0.0465
29	v5a11	12.04	15	0.6767
30	v5a14	12.61	10	0.2457
31	v5a18	9.76	15	0.8349
32	v5b2	9.39	15	0.8565
33	v5b5	11.03	13	0.6096
34	v5b6	18.02	14	0.2052
35	v5b8	14.69	13	0.3291
36	v5b13	13.05	9	0.1601

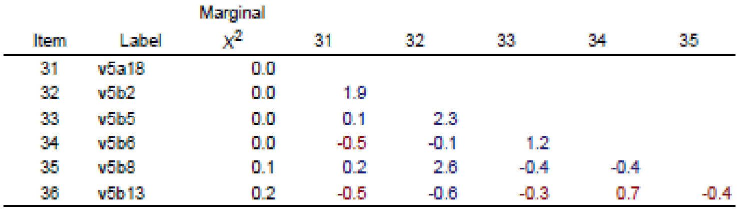
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

Item	Label	Marginal X^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.0										
2	m3	0.0	14.3									
3	m1	0.0	5.5	-0.1								
4	m13	0.0	15.7	-0.5	0.7							
5	m18	0.0	-0.6	2.6	3.9	7.2						
6	m9	0.0	0.5	-0.7	1.4	2.2	0.7					
7	m4	0.2	9.7	4.9	-0.6	-0.6	5.1	-0.5				
8	m16	0.1	13.4	-0.2	1.0	-0.2	1.8	-0.6	0.0			
9	m14	0.3	3.8	1.9	7.3	-0.2	8.2	-0.4	0.8	-0.2		
10	m6	0.0	8.8	2.6	-0.4	1.7	8.3	1.1	0.0	0.8	-0.2	
11	m11	0.2	3.8	-0.5	-0.3	4.3	3.6	0.3	0.2	3.1	1.4	0.2
12	m7	0.3	7.8	2.3	-0.3	-0.4	5.2	-0.5	3.9	-0.4	-0.3	0.3
13	m5	0.0	3.3	-0.6	-0.3	4.6	1.6	-0.3	-0.5	-0.3	5.8	-0.6
14	m2	0.0	3.1	3.7	-0.6	0.7	0.4	0.4	7.1	5.3	7.0	8.5
15	m10	0.0	-0.3	0.7	-0.7	-0.2	-0.1	-0.2	3.3	4.5	7.4	8.5
16	m15	0.0	2.6	0.0	1.3	2.8	-0.2	-0.6	2.4	-0.3	3.6	0.1
17	m8	0.2	6.1	-0.5	0.5	-0.0	5.0	-0.5	-0.5	0.1	1.6	1.3
18	m17	0.0	2.6	1.8	-0.3	-0.6	3.1	-0.5	-0.6	0.9	-0.3	0.2
19	v3a1	0.0	7.8	1.0	-0.6	0.1	2.4	0.1	-0.6	-0.4	-0.0	2.8
20	v3a3	0.1	16.9	2.3	-0.3	0.7	3.2	-0.2	0.1	-0.6	5.8	0.2
21	v3a9	0.1	0.8	0.2	-0.4	1.3	1.5	1.4	-0.6	2.3	9.5	1.3
22	v3a17	0.1	12.3	-0.6	-0.4	-0.1	0.7	0.1	3.0	-0.4	0.0	0.4
23	v3b4	0.1	12.8	-0.6	-0.5	3.0	8.6	-0.3	3.8	1.2	-0.3	-0.4
24	v3b12	0.2	12.9	2.6	0.9	5.3	0.7	2.4	2.8	1.0	-0.2	5.3
25	v3b15	0.1	-0.0	6.3	4.9	6.1	2.7	6.3	5.0	1.3	0.1	3.0
26	v3b16	0.0	14.0	-0.6	0.6	0.6	0.2	1.9	-0.7	1.4	2.9	-0.4
27	v5a7	0.0	8.6	3.5	-0.4	-0.1	2.4	-0.4	-0.6	-0.6	-0.3	-0.6
28	v5a10	0.0	7.5	-0.3	-0.4	0.0	3.4	-0.4	-0.3	-0.5	0.6	0.8
29	v5a11	0.0	12.9	-0.6	1.1	-0.5	4.4	1.8	1.8	-0.1	-0.2	2.5
30	v5a14	0.2	0.1	4.0	0.5	-0.3	2.1	2.2	-0.5	-0.4	0.4	4.0
31	v5a18	0.0	0.2	0.4	0.9	-0.4	6.1	-0.1	0.1	3.0	-0.5	1.6
32	v5b2	0.0	3.5	-0.2	-0.1	1.4	3.7	1.2	-0.6	-0.5	1.9	0.7
33	v5b5	0.0	10.8	-0.1	0.1	1.0	0.3	8.1	-0.1	2.7	4.2	-0.1
34	v5b6	0.0	11.4	-0.6	1.5	0.6	2.6	-0.1	-0.5	-0.2	0.3	0.0
35	v5b8	0.1	3.2	-0.5	0.4	0.3	1.4	0.1	0.2	0.7	-0.5	1.0
36	v5b13	0.2	6.1	1.1	-0.0	1.8	0.9	0.6	1.5	-0.4	0.1	4.8

Item	Label	Marginal X^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.2										
12	m7	0.3	0.2									
13	m5	0.0	2.0	-0.4								
14	m2	0.0	5.1	5.0	0.1							
15	m10	0.0	1.2	0.5	3.1	1.5						
16	m15	0.0	5.1	4.2	-0.1	-0.3	0.0					
17	m8	0.2	3.0	0.4	0.4	3.0	2.9	1.8				
18	m17	0.0	-0.2	0.2	-0.6	6.5	2.0	1.4	-0.4			
19	v3a1	0.0	2.1	-0.6	2.0	-0.2	-0.6	-0.4	-0.5	-0.7		
20	v3a3	0.1	0.7	-0.6	-0.2	1.5	3.8	-0.4	0.7	1.2	1.8	
21	v3a9	0.1	4.1	0.8	1.3	0.2	1.8	0.2	0.5	0.8	-0.1	0.1
22	v3a17	0.1	2.9	0.4	0.0	7.2	0.4	-0.5	2.3	-0.4	-0.4	-0.6
23	v3b4	0.1	1.6	0.9	0.3	2.0	2.0	-0.2	-0.5	-0.6	-0.1	-0.1
24	v3b12	0.2	1.8	2.6	1.7	0.9	1.4	1.1	0.4	5.2	-0.1	2.1
25	v3b15	0.1	0.9	6.5	13.5	1.5	1.5	5.0	1.1	4.1	4.0	5.6
26	v3b16	0.0	7.7	-0.1	2.7	0.4	2.1	-0.2	-0.5	0.3	-0.4	1.5
27	v5a7	0.0	5.3	1.1	0.1	1.9	0.4	0.3	-0.3	-0.4	3.7	-0.2
28	v5a10	0.0	0.2	-0.4	1.5	5.8	0.1	-0.2	0.4	2.0	0.9	1.7
29	v5a11	0.0	-0.2	0.4	2.0	2.7	5.8	2.9	1.4	-0.1	0.4	-0.6
30	v5a14	0.2	2.0	-0.2	5.1	0.7	1.2	2.6	0.9	-0.4	0.1	7.4
31	v5a18	0.0	0.1	-0.4	3.8	2.4	3.3	3.9	-0.3	-0.6	1.4	4.4
32	v5b2	0.0	-0.1	-0.6	0.3	-0.3	1.4	2.0	0.1	-0.5	5.2	2.1
33	v5b5	0.0	-0.7	-0.6	2.3	3.9	0.8	1.6	3.8	-0.7	1.1	0.6
34	v5b6	0.0	3.1	-0.3	4.0	1.7	5.6	-0.1	0.2	-0.6	0.2	-0.4
35	v5b8	0.1	7.9	1.0	-0.1	2.5	-0.3	0.2	6.1	2.2	2.5	0.4
36	v5b13	0.2	-0.4	1.3	1.7	1.3	-0.6	-0.0	0.5	0.8	0.4	-0.5



Item	Label	Marginal χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.1										
22	v3a17	0.1	0.8									
23	v3b4	0.1	3.0	-0.5								
24	v3b12	0.2	0.3	-0.3	4.5							
25	v3b15	0.1	5.6	5.7	10.9	0.6						
26	v3b16	0.0	1.8	-0.4	0.2	2.0	6.8					
27	v5a7	0.0	-0.7	-0.4	-0.5	1.7	7.8	-0.4				
28	v5a10	0.0	0.9	0.1	0.4	1.2	3.8	1.4	-0.7			
29	v5a11	0.0	0.2	-0.4	1.0	-0.5	1.0	-0.2	-0.3	0.5		
30	v5a14	0.2	3.8	1.8	-0.5	-0.4	4.8	0.3	-0.4	0.4	2.9	
31	v5a18	0.0	3.4	0.2	5.6	5.2	8.2	1.0	-0.3	1.0	-0.3	0.3
32	v5b2	0.0	0.3	1.1	-0.7	-0.5	0.5	-0.3	2.5	0.3	-0.6	-0.6
33	v5b5	0.0	-0.7	-0.6	-0.3	2.0	1.8	-0.4	-0.2	-0.6	-0.5	1.9
34	v5b6	0.0	-0.4	-0.1	-0.5	2.4	8.8	-0.5	2.6	-0.4	-0.3	1.4
35	v5b8	0.1	-0.7	4.0	-0.5	-0.4	0.7	-0.6	-0.5	1.5	0.7	-0.5
36	v5b13	0.2	-0.0	-0.3	-0.3	-0.4	7.4	-0.5	0.5	-0.3	-0.5	-0.1



Item	Label	Marginal χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	1.9				
33	v5b5	0.0	0.1	2.3			
34	v5b6	0.0	-0.5	-0.1	1.2		
35	v5b8	0.1	0.2	2.6	-0.4	-0.4	
36	v5b13	0.2	-0.5	-0.6	-0.3	0.7	-0.4

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	6178.21
Akaike Information Criterion (AIC):	6250.21
Bayesian Information Criterion (BIC):	6357.87

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	36
Number of Dimensions	1

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - All Items; 1PL

Item	*Label*		*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e.*
1	m12	37	0.58	0.05	1	0.1	0.17	-0.18	0.29
2	m3	37	0.58	0.05	2	0.16	0.17	-0.28	0.3
3	m1	37	0.58	0.05	3	-0.05	0.17	0.09	0.3
4	m13	37	0.58	0.05	4	0.31	0.17	-0.54	0.3
5	m18	37	0.58	0.05	5	0.06	0.17	-0.1	0.3
6	m9	37	0.58	0.05	6	-0.22	0.17	0.38	0.3
7	m4	37	0.58	0.05	7	1.13	0.19	-1.95	0.37
8	m16	37	0.58	0.05	8	0.63	0.18	-1.09	0.32
9	m14	37	0.58	0.05	9	1.91	0.24	-3.28	0.5
10	m6	37	0.58	0.05	10	0.06	0.17	-0.1	0.3
11	m11	37	0.58	0.05	11	1.33	0.2	-2.29	0.4
12	m7	37	0.58	0.05	12	1.44	0.21	-2.48	0.42
13	m5	37	0.58	0.05	13	-0.29	0.17	0.5	0.3
14	m2	37	0.58	0.05	14	-0.13	0.17	0.23	0.3
15	m10	37	0.58	0.05	15	-0.25	0.17	0.43	0.3
16	m15	37	0.58	0.05	16	0.12	0.17	-0.2	0.3
17	m8	37	0.58	0.05	17	1.29	0.2	-2.22	0.39
18	m17	37	0.58	0.05	18	0.55	0.18	-0.95	0.31
19	v3a1	37	0.58	0.05	19	0.68	0.19	-1.18	0.33
20	v3a3	37	0.58	0.05	20	-0.31	0.18	0.54	0.31
21	v3a9	37	0.58	0.05	21	-0.42	0.18	0.72	0.32
22	v3a17	37	0.58	0.05	22	1.16	0.2	-1.99	0.38
23	v3b4	37	0.58	0.05	23	0.81	0.19	-1.4	0.35
24	v3b12	37	0.58	0.05	24	1.6	0.23	-2.76	0.46
25	v3b15	37	0.58	0.05	25	-0.53	0.19	0.91	0.33
26	v3b16	37	0.58	0.05	26	-0.11	0.18	0.19	0.31
27	v5a7	37	0.58	0.05	27	0.2	0.19	-0.34	0.32
28	v5a10	37	0.58	0.05	28	0.09	0.18	-0.16	0.32
29	v5a11	37	0.58	0.05	29	-0.11	0.19	0.19	0.32
30	v5a14	37	0.58	0.05	30	1.41	0.22	-2.43	0.43
31	v5a18	37	0.58	0.05	31	0.13	0.18	-0.22	0.32
32	v5b2	37	0.58	0.05	32	0.28	0.19	-0.48	0.33
33	v5b5	37	0.58	0.05	33	0.64	0.2	-1.09	0.35
34	v5b6	37	0.58	0.05	34	0.38	0.19	-0.66	0.33
35	v5b8	37	0.58	0.05	35	1.05	0.21	-1.8	0.39
36	v5b13	37	0.58	0.05	36	1.47	0.23	-2.54	0.45

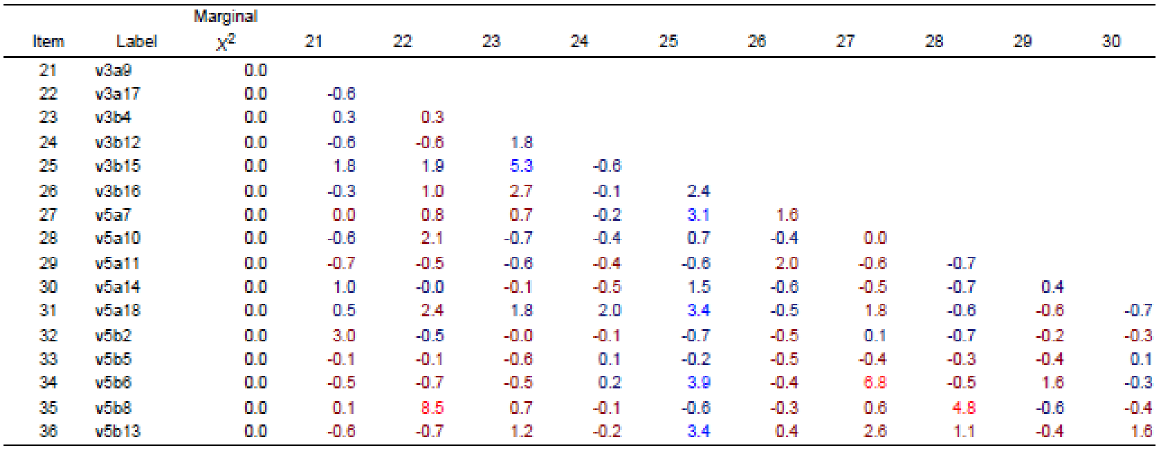
S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	Probability
1	m12	45.62	14	0.0001
2	m3	21.04	14	0.1003
3	m1	13	13	0.4494
4	m13	14.61	14	0.4067
5	m18	28.28	14	0.013
6	m9	12.06	13	0.5242
7	m4	16.54	11	0.122
8	m16	29.22	14	0.0097
9	m14	11.35	8	0.1821
10	m6	19.09	15	0.2093
11	m11	19.04	13	0.1214
12	m7	14.67	11	0.1975
13	m5	10.29	13	0.6713
14	m2	21.02	14	0.1008
15	m10	22.14	14	0.0757
16	m15	13.06	14	0.5235
17	m8	14.51	12	0.2688
18	m17	12.91	13	0.4566
19	v3a1	11.32	12	0.5032
20	v3a3	12.87	13	0.4596
21	v3a9	15.21	13	0.2937
22	v3a17	13.01	10	0.2223
23	v3b4	10.93	12	0.536
24	v3b12	17.36	10	0.0665
25	v3b15	39.24	15	0.0006
26	v3b16	11.47	13	0.5724
27	v5a7	18.86	13	0.1272
28	v5a10	24.86	14	0.0358
29	v5a11	11.48	14	0.6489
30	v5a14	9.65	10	0.4729
31	v5a18	8.68	14	0.8513
32	v5b2	10.53	15	0.7856
33	v5b5	11.61	13	0.561
34	v5b6	20.08	14	0.1273
35	v5b8	14.93	11	0.1854
36	v5b13	15.88	9	0.0692

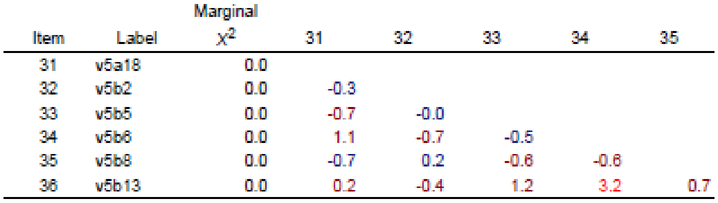
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

Item	Label	Marginal χ^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.0										
2	m3	0.0	7.1									
3	m1	0.0	1.3	2.5								
4	m13	0.0	8.2	1.5	-0.7							
5	m18	0.0	-0.3	-0.1	0.5	2.4						
6	m9	0.0	-0.7	0.5	5.4	-0.3	-0.7					
7	m4	0.0	4.5	10.1	-0.1	0.0	1.4	-0.3				
8	m16	0.0	6.8	2.1	-0.6	-0.6	-0.4	-0.0	1.9			
9	m14	0.0	1.1	-0.0	3.5	-0.6	4.2	-0.5	3.0	1.0		
10	m6	0.0	3.3	-0.1	-0.5	-0.5	3.1	-0.6	2.3	4.0	-0.7	
11	m11	0.0	0.7	-0.3	1.3	1.1	0.6	-0.7	-0.7	0.5	-0.1	-0.7
12	m7	0.0	3.5	5.9	1.2	-0.4	1.7	-0.3	8.5	0.9	-0.6	-0.7
13	m5	0.0	0.2	-0.1	1.7	0.9	-0.5	-0.5	0.5	-0.6	2.3	1.2
14	m2	0.0	0.1	0.4	-0.3	-0.7	-0.7	-0.7	2.6	1.4	3.3	3.2
15	m10	0.0	-0.6	-0.7	0.4	-0.6	-0.6	-0.6	0.4	0.9	3.5	3.2
16	m15	0.0	-0.1	-0.7	-0.6	-0.0	-0.6	-0.2	-0.0	-0.4	1.0	-0.7
17	m8	0.0	2.1	-0.1	-0.7	-0.7	1.5	-0.1	0.2	2.2	4.5	-0.5
18	m17	0.0	-0.1	-0.4	-0.6	0.0	0.2	-0.3	0.1	4.2	-0.6	-0.7
19	v3a1	0.0	3.1	-0.5	0.5	-0.6	-0.1	-0.3	-0.1	-0.5	-0.7	0.1
20	v3a3	0.0	9.3	6.4	-0.3	-0.6	0.2	-0.1	-0.5	0.9	2.4	-0.7
21	v3a9	0.0	-0.6	-0.6	0.2	-0.4	-0.5	-0.3	0.4	-0.1	5.1	-0.5
22	v3a17	0.0	6.7	0.1	0.1	-0.6	-0.6	1.3	7.8	0.9	1.3	-0.7
23	v3b4	0.0	6.8	0.4	-0.3	0.4	3.7	-0.2	8.9	4.4	-0.4	-0.5
24	v3b12	0.0	7.8	0.3	-0.5	2.1	-0.6	0.2	0.8	-0.4	-0.5	2.0
25	v3b15	0.0	-0.6	2.1	1.3	2.0	0.1	2.2	1.6	-0.5	-0.6	0.2
26	v3b16	0.0	7.2	1.1	-0.6	-0.5	-0.7	-0.2	-0.1	5.0	0.6	-0.4
27	v5a7	0.0	3.6	8.2	0.6	2.1	-0.0	0.0	1.2	0.5	-0.6	-0.3
28	v5a10	0.0	2.8	-0.4	0.3	-0.3	0.4	0.7	-0.3	1.1	-0.6	-0.6
29	v5a11	0.0	6.6	-0.0	-0.4	0.5	1.0	-0.2	0.0	-0.4	-0.7	-0.0
30	v5a14	0.0	-0.6	1.1	-0.4	0.2	0.1	0.1	0.2	-0.3	2.0	1.1
31	v5a18	0.0	-0.6	-0.6	-0.4	0.1	11.6	-0.4	-0.4	0.4	-0.4	-0.4
32	v5b2	0.0	0.6	-0.5	1.1	-0.1	0.7	-0.2	0.6	0.1	0.1	-0.6
33	v5b5	0.0	5.5	-0.5	0.2	-0.1	-0.5	3.5	2.4	0.4	1.8	-0.7
34	v5b6	0.0	5.7	-0.1	-0.1	3.4	0.2	0.1	-0.1	1.8	-0.6	2.4
35	v5b8	0.0	0.6	0.4	-0.2	-0.2	-0.2	1.1	2.9	3.4	0.2	-0.5
36	v5b13	0.0	2.7	3.8	0.2	4.8	-0.3	2.0	5.1	0.9	-0.6	1.7

Item	Label	Marginal χ^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.0										
12	m7	0.0	-0.7									
13	m5	0.0	-0.2	0.7								
14	m2	0.0	1.5	1.5	-0.7							
15	m10	0.0	-0.5	-0.7	0.1	-0.5						
16	m15	0.0	1.5	1.1	-0.6	-0.5	-0.7					
17	m8	0.0	0.6	2.6	-0.7	0.3	0.2	-0.3				
18	m17	0.0	-0.6	2.4	-0.3	2.0	-0.3	-0.5	-0.5			
19	v3a1	0.0	0.1	-0.0	-0.1	-0.6	0.7	-0.3	1.0	-0.0		
20	v3a3	0.0	-0.5	0.3	1.1	-0.4	0.5	-0.2	-0.6	-0.5	-0.4	
21	v3a9	0.0	1.1	-0.5	-0.1	-0.5	-0.4	-0.4	-0.6	-0.6	-0.7	-0.7
22	v3a17	0.0	0.8	-0.6	0.9	2.9	-0.7	0.0	6.5	-0.6	-0.6	0.4
23	v3b4	0.0	-0.0	3.2	1.2	-0.2	-0.2	0.6	0.4	0.4	1.4	1.8
24	v3b12	0.0	0.3	0.7	0.1	-0.5	-0.3	-0.1	-0.6	2.2	-0.5	0.0
25	v3b15	0.0	3.8	2.6	7.4	-0.4	-0.4	1.4	-0.5	0.8	0.8	1.7
26	v3b16	0.0	3.5	-0.6	0.3	-0.7	-0.2	1.0	0.1	-0.7	0.4	5.2
27	v5a7	0.0	2.1	3.8	0.1	-0.3	-0.7	-0.2	0.2	1.6	0.8	1.6
28	v5a10	0.0	-0.3	-0.5	-0.1	1.7	2.7	0.4	-0.6	-0.2	-0.3	-0.4
29	v5a11	0.0	2.2	-0.7	0.0	0.1	1.8	0.4	-0.4	-0.7	-0.2	0.6
30	v5a14	0.0	0.4	0.5	1.8	-0.6	-0.4	0.5	-0.4	0.8	1.0	3.2
31	v5a18	0.0	-0.3	-0.5	0.8	-0.1	0.3	0.9	-0.1	-0.3	-0.2	1.0
32	v5b2	0.0	-0.5	0.6	-0.3	1.7	-0.4	0.1	-0.7	-0.4	9.8	-0.2
33	v5b5	0.0	0.5	-0.4	5.5	0.8	-0.5	-0.1	1.1	-0.1	-0.3	-0.6
34	v5b6	0.0	0.8	-0.6	0.9	-0.3	1.7	0.8	-0.7	0.8	-0.3	1.2
35	v5b8	0.0	4.5	-0.4	0.1	0.1	-0.6	-0.1	11.5	0.0	0.3	-0.7
36	v5b13	0.0	-0.4	4.1	-0.1	-0.3	-0.1	0.5	-0.6	3.3	1.7	-0.0



Item	Label	Marginal χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.0										
22	v3a17	0.0	-0.6									
23	v3b4	0.0	0.3	0.3								
24	v3b12	0.0	-0.6	-0.6	1.8							
25	v3b15	0.0	1.8	1.9	5.3	-0.6						
26	v3b16	0.0	-0.3	1.0	2.7	-0.1	2.4					
27	v5a7	0.0	0.0	0.8	0.7	-0.2	3.1	1.6				
28	v5a10	0.0	-0.6	2.1	-0.7	-0.4	0.7	-0.4	0.0			
29	v5a11	0.0	-0.7	-0.5	-0.6	-0.4	-0.6	2.0	-0.6	-0.7		
30	v5a14	0.0	1.0	-0.0	-0.1	-0.5	1.5	-0.6	-0.5	-0.7	0.4	
31	v5a18	0.0	0.5	2.4	1.8	2.0	3.4	-0.5	1.8	-0.6	-0.6	-0.7
32	v5b2	0.0	3.0	-0.5	-0.0	-0.1	-0.7	-0.5	0.1	-0.7	-0.2	-0.3
33	v5b5	0.0	-0.1	-0.1	-0.6	0.1	-0.2	-0.5	-0.4	-0.3	-0.4	0.1
34	v5b6	0.0	-0.5	-0.7	-0.5	0.2	3.9	-0.4	6.8	-0.5	1.6	-0.3
35	v5b8	0.0	0.1	8.5	0.7	-0.1	-0.6	-0.3	0.6	4.8	-0.6	-0.4
36	v5b13	0.0	-0.6	-0.7	1.2	-0.2	3.4	0.4	2.6	1.1	-0.4	1.6



Item	Label	Marginal χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	-0.3				
33	v5b5	0.0	-0.7	-0.0			
34	v5b6	0.0	1.1	-0.7	-0.5		
35	v5b8	0.0	-0.7	0.2	-0.6	-0.6	
36	v5b13	0.0	0.2	-0.4	1.2	3.2	0.7

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	6139.37
Akaike Information Criterion (AIC):	6213.37
Bayesian Information Criterion (BIC):	6324.02

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	36
Number of Dimensions	1

IRTPRO Version 4.2

Output generated by IRTPRO estimation engine Version 5.20 (64-bit)

Project: Combined Sample - All Items; 2PL

Item	***Label***		***a***	***s.e.***		***c***	***s.e.***	***b***	***s.e.***
1	m12	2	-0.76	0.23	1	0.1	0.18	0.13	0.25
2	m3	4	1.44	0.34	3	0.25	0.23	-0.17	0.16
3	m1	6	0.81	0.24	5	-0.04	0.18	0.05	0.23
4	m13	8	0.67	0.22	7	0.32	0.18	-0.48	0.29
5	m18	10	-0.1	0.19	9	0.06	0.16	0.56	1.99
6	m9	12	0.78	0.23	11	-0.22	0.18	0.29	0.25
7	m4	14	2.17	0.57	13	1.84	0.46	-0.85	0.15
8	m16	16	1.35	0.33	15	0.81	0.25	-0.6	0.18
9	m14	18	0.44	0.29	17	1.86	0.25	-4.26	2.71
10	m6	20	0.49	0.2	19	0.06	0.17	-0.12	0.35
11	m11	22	0.1	0.22	21	1.24	0.2	-12.67	28.61
12	m7	24	1.69	0.47	23	1.97	0.4	-1.17	0.21
13	m5	26	0.63	0.22	25	-0.29	0.18	0.46	0.32
14	m2	28	-0.02	0.19	27	-0.12	0.16	-5.78	50.18
15	m10	30	0.14	0.19	29	-0.23	0.17	1.64	2.44
16	m15	32	0.41	0.2	31	0.11	0.17	-0.28	0.43
17	m8	34	0.99	0.31	33	1.43	0.27	-1.45	0.38
18	m17	36	0.85	0.25	35	0.6	0.19	-0.71	0.26
19	v3a1	38	0.69	0.24	37	0.72	0.2	-1.04	0.39
20	v3a3	40	0.87	0.25	39	-0.3	0.19	0.35	0.24
21	v3a9	42	0.35	0.2	41	-0.4	0.18	1.16	0.86
22	v3a17	44	1.39	0.4	43	1.5	0.34	-1.08	0.23
23	v3b4	46	1.44	0.38	45	1.1	0.3	-0.76	0.19
24	v3b12	48	0.05	0.25	47	1.5	0.22	-30.28	154.31
25	v3b15	50	-0.47	0.22	49	-0.53	0.19	-1.12	0.6
26	v3b16	52	0.94	0.26	51	-0.09	0.2	0.09	0.22
27	v5a7	54	1.45	0.35	53	0.34	0.25	-0.24	0.16
28	v5a10	56	0.66	0.23	55	0.12	0.19	-0.18	0.28
29	v5a11	58	0.52	0.22	57	-0.1	0.18	0.19	0.37
30	v5a14	60	0.4	0.26	59	1.37	0.23	-3.42	2.14
31	v5a18	62	0.36	0.21	61	0.13	0.18	-0.35	0.51
32	v5b2	64	0.53	0.23	63	0.29	0.19	-0.55	0.39
33	v5b5	66	0.56	0.24	65	0.65	0.2	-1.16	0.53
34	v5b6	68	0.94	0.28	67	0.47	0.22	-0.5	0.23
35	v5b8	70	1.07	0.34	69	1.24	0.29	-1.16	0.3
36	v5b13	72	2.05	0.75	71	2.34	0.68	-1.14	0.19

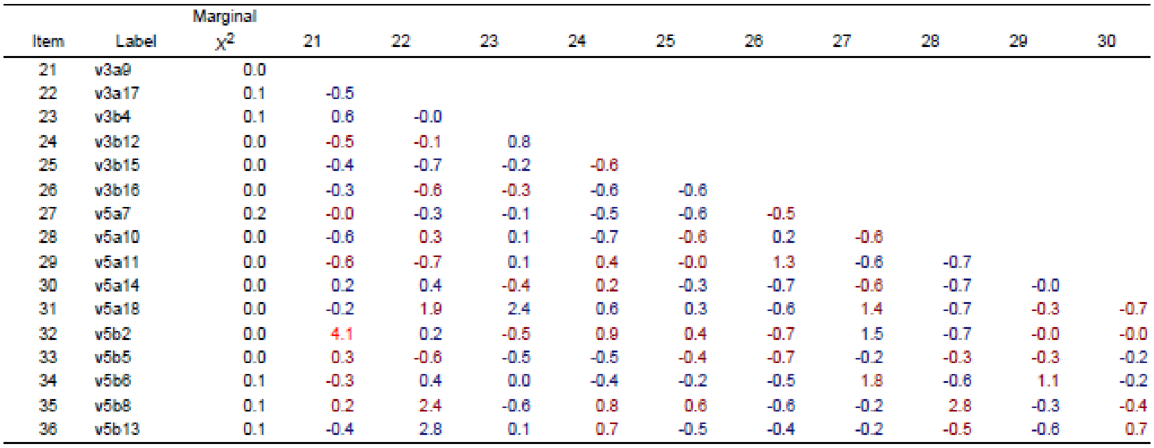
S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	Probability
1	m12	12.11	13	0.5197
2	m3	17.33	9	0.0437
3	m1	12.23	13	0.5103
4	m13	15.73	15	0.4016
5	m18	23.88	17	0.1224
6	m9	11.56	13	0.5651
7	m4	9.01	9	0.4381
8	m16	24.86	11	0.0095
9	m14	9.19	8	0.3284
10	m6	17.5	14	0.2301
11	m11	14.19	12	0.2877
12	m7	12.91	8	0.1147
13	m5	11.04	12	0.5264
14	m2	15.46	14	0.3493
15	m10	20.68	16	0.1907
16	m15	13.5	15	0.5652
17	m8	13.4	11	0.2671
18	m17	12.66	12	0.396
19	v3a1	11.16	12	0.5166
20	v3a3	18.73	14	0.1751
21	v3a9	16.84	14	0.2638
22	v3a17	12.25	9	0.1989
23	v3b4	11.67	9	0.2318
24	v3b12	13.33	11	0.2716
25	v3b15	16.97	15	0.3228
26	v3b16	12.77	14	0.5461
27	v5a7	15.18	10	0.1252
28	v5a10	24.91	14	0.0354
29	v5a11	11.73	15	0.7003
30	v5a14	8.19	11	0.6975
31	v5a18	11.07	16	0.8056
32	v5b2	11.03	15	0.7511
33	v5b5	11.63	13	0.5596
34	v5b6	19.2	11	0.0575
35	v5b8	11.64	10	0.3114
36	v5b13	7.18	7	0.4123

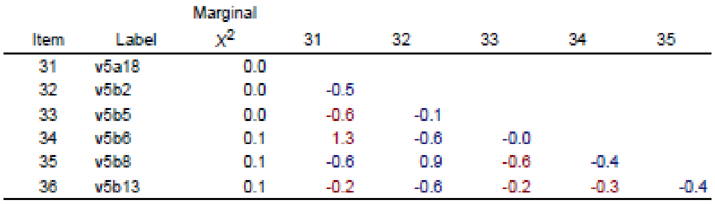
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

		Marginal										
Item	Label	χ^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.0										
2	m3	0.0	-0.6									
3	m1	0.0	-0.4	-0.2								
4	m13	0.0	0.8	-0.3	-0.5							
5	m18	0.0	0.8	-0.6	-0.7	0.1						
6	m9	0.0	2.8	-0.7	3.2	0.3	-0.1					
7	m4	0.0	-0.5	0.0	-0.2	-0.6	-0.4	-0.1				
8	m16	0.0	-0.6	-0.6	1.2	-0.5	-0.5	-0.7	-0.0			
9	m14	0.0	-0.6	0.5	3.5	-0.6	1.9	-0.5	1.5	0.5		
10	m6	0.0	-0.5	0.9	-0.5	-0.5	0.5	-0.6	0.4	2.4	-0.6	
11	m11	0.0	-0.5	0.4	2.8	0.0	-0.4	-0.5	-0.6	-0.2	-0.5	-0.4
12	m7	0.0	-0.6	-0.3	-0.7	-0.5	-0.2	-0.5	-0.6	-0.2	-0.7	-0.3
13	m5	0.0	-0.1	-0.7	0.9	1.3	-0.6	-0.7	-0.7	-0.6	2.0	1.3
14	m2	0.0	-0.7	-0.6	1.3	-0.3	-0.2	-0.1	0.5	-0.2	1.4	0.8
15	m10	0.0	1.1	-0.6	1.7	0.1	0.4	-0.0	-0.1	0.1	1.8	1.3
16	m15	0.0	-0.3	-0.7	-0.6	-0.2	0.6	-0.1	1.0	-0.6	0.4	-0.5
17	m8	0.0	-0.7	-0.4	-0.1	-0.6	-0.3	-0.7	-0.2	-0.5	4.0	-0.2
18	m17	0.0	0.7	2.6	-0.6	-0.5	-0.7	-0.7	-0.3	0.5	-0.6	-0.7
19	v3a1	0.0	-0.5	0.9	-0.1	-0.6	-0.7	-0.4	-0.1	-0.4	-0.6	0.1
20	v3a3	0.1	0.9	1.7	-0.5	-0.2	-0.7	-0.4	2.2	-0.6	2.5	-0.7
21	v3a9	0.0	0.6	-0.6	0.4	-0.6	-0.6	-0.3	-0.0	0.1	3.8	-0.6
22	v3a17	0.1	-0.3	0.4	-0.4	-0.3	-0.3	-0.1	-0.3	-0.3	0.7	-0.3
23	v3b4	0.1	-0.4	0.2	-0.4	2.7	0.8	-0.3	-0.2	-0.4	-0.6	-0.6
24	v3b12	0.0	4.6	-0.4	-0.6	0.8	-0.6	-0.4	0.1	-0.6	0.1	0.6
25	v3b15	0.0	-0.6	-0.7	-0.6	-0.4	-0.6	-0.5	-0.2	0.8	0.2	-0.6
26	v3b16	0.0	0.0	-0.6	0.0	-0.1	0.2	0.9	0.5	0.8	0.7	-0.5
27	v5a7	0.2	-0.4	0.7	-0.3	0.3	-0.4	-0.3	1.6	0.4	-0.6	-0.6
28	v5a10	0.0	-0.5	-0.3	-0.1	-0.4	-0.6	0.2	0.5	-0.2	-0.6	-0.6
29	v5a11	0.0	0.9	-0.4	-0.2	0.5	-0.4	-0.1	1.8	-0.3	-0.6	-0.2
30	v5a14	0.0	0.7	1.8	-0.4	0.4	-0.6	0.0	-0.1	-0.3	2.9	0.5
31	v5a18	0.0	1.3	-0.4	-0.4	0.4	17.5	-0.2	-0.2	0.8	-0.0	-0.6
32	v5b2	0.0	-0.2	-0.4	0.9	-0.1	-0.4	-0.1	-0.2	-0.3	-0.1	-0.7
33	v5b5	0.0	0.7	-0.3	0.1	-0.1	0.1	4.2	0.4	1.8	1.5	-0.6
34	v5b6	0.1	-0.0	-0.2	0.8	2.0	-0.4	-0.2	1.0	-0.2	-0.5	1.8
35	v5b8	0.1	0.4	-0.3	0.1	-0.1	-0.3	0.1	-0.3	0.1	0.1	-0.1
36	v5b13	0.1	-0.1	-0.5	0.1	1.5	-0.2	0.0	0.1	0.7	-0.3	4.2

		Marginal										
Item	Label	χ^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.0										
12	m7	0.0	-0.6									
13	m5	0.0	-0.7	-0.6								
14	m2	0.0	0.0	-0.0	0.2							
15	m10	0.0	-0.7	-0.7	-0.6	-0.6						
16	m15	0.0	0.2	2.2	-0.5	0.7	-0.1					
17	m8	0.0	-0.1	-0.5	-0.5	-0.6	-0.4	-0.2				
18	m17	0.0	-0.1	-0.5	-0.6	0.1	-0.7	-0.5	-0.6			
19	v3a1	0.0	-0.4	-0.6	0.2	0.3	2.0	-0.1	-0.0	-0.5		
20	v3a3	0.1	-0.5	-0.6	0.5	-0.4	-0.2	-0.2	0.4	0.4	0.4	
21	v3a9	0.0	0.0	-0.3	-0.1	0.2	-0.7	-0.1	-0.6	-0.7	-0.5	-0.6
22	v3a17	0.1	0.1	3.9	-0.0	0.8	-0.6	-0.2	1.4	-0.1	-0.5	-0.6
23	v3b4	0.1	-0.4	-0.6	0.2	-0.7	-0.5	0.2	-0.6	-0.7	-0.2	-0.4
24	v3b12	0.0	-0.3	0.0	-0.0	-0.7	-0.7	-0.2	-0.7	0.8	0.0	-0.6
25	v3b15	0.0	6.9	-0.7	2.3	-0.6	-0.5	-0.2	0.1	-0.7	-0.7	-0.7
26	v3b16	0.0	1.8	0.5	1.0	-0.2	-0.6	0.9	-0.7	-0.2	-0.3	2.2
27	v5a7	0.2	1.1	-0.6	-0.1	-0.5	-0.5	-0.1	-0.4	-0.4	3.3	-0.4
28	v5a10	0.0	-0.1	-0.5	-0.0	0.0	4.7	0.6	-0.4	0.4	-0.3	0.3
29	v5a11	0.0	4.1	-0.2	0.0	-0.7	0.4	0.1	-0.1	-0.7	-0.3	0.3
30	v5a14	0.0	-0.1	-0.0	1.4	-0.6	-0.7	0.1	-0.4	0.8	1.2	3.2
31	v5a18	0.0	0.0	-0.7	0.5	-0.7	-0.5	0.3	-0.1	-0.2	-0.4	0.9
32	v5b2	0.0	-0.2	-0.4	-0.3	4.3	-0.6	-0.1	-0.6	-0.6	9.5	0.2
33	v5b5	0.0	1.8	-0.7	5.4	-0.4	-0.5	-0.2	2.0	-0.3	-0.3	-0.4
34	v5b6	0.1	0.0	0.6	2.0	-0.6	0.7	0.7	0.1	-0.3	-0.3	-0.1
35	v5b8	0.1	3.1	3.0	-0.2	-0.6	-0.2	-0.1	6.0	2.0	1.6	0.3
36	v5b13	0.1	-0.0	-0.6	1.8	-0.6	0.3	0.0	2.5	-0.2	-0.1	-0.4



Item	Label	Marginal χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.0										
22	v3a17	0.1	-0.5									
23	v3b4	0.1	0.6	-0.0								
24	v3b12	0.0	-0.5	-0.1	0.8							
25	v3b15	0.0	-0.4	-0.7	-0.2	-0.6						
26	v3b16	0.0	-0.3	-0.6	-0.3	-0.6	-0.6					
27	v5a7	0.2	-0.0	-0.3	-0.1	-0.5	-0.6	-0.5				
28	v5a10	0.0	-0.6	0.3	0.1	-0.7	-0.6	0.2	-0.6			
29	v5a11	0.0	-0.6	-0.7	0.1	0.4	-0.0	1.3	-0.6	-0.7		
30	v5a14	0.0	0.2	0.4	-0.4	0.2	-0.3	-0.7	-0.6	-0.7	-0.0	
31	v5a18	0.0	-0.2	1.9	2.4	0.6	0.3	-0.6	1.4	-0.7	-0.3	-0.7
32	v5b2	0.0	4.1	0.2	-0.5	0.9	0.4	-0.7	1.5	-0.7	-0.0	-0.0
33	v5b5	0.0	0.3	-0.6	-0.5	-0.5	-0.4	-0.7	-0.2	-0.3	-0.3	-0.2
34	v5b6	0.1	-0.3	0.4	0.0	-0.4	-0.2	-0.5	1.8	-0.6	1.1	-0.2
35	v5b8	0.1	0.2	2.4	-0.6	0.8	0.6	-0.6	-0.2	2.8	-0.3	-0.4
36	v5b13	0.1	-0.4	2.8	0.1	0.7	-0.5	-0.4	-0.2	-0.5	-0.6	0.7



Item	Label	Marginal χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	-0.5				
33	v5b5	0.0	-0.6	-0.1			
34	v5b6	0.1	1.3	-0.6	-0.0		
35	v5b8	0.1	-0.6	0.9	-0.6	-0.4	
36	v5b13	0.1	-0.2	-0.6	-0.2	-0.3	-0.4

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	5954.15
Akaike Information Criterion (AIC):	6098.15
Bayesian Information Criterion (BIC):	6313.47

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	36
Number of Dimensions	1

IRTPRO Version 4.2

Output generated by IRTPRO estimation engine Version 5.20 (64-bit)

Project: Combined Sample - All Items; 3PL

Item	*Label*		*a*	*s.e.*		*c*	*s.e.*	*b*	*s.e.*		*logit g*	*s.e.*	*g*	*s.e.*
1	m12	3	0.09	0.05	2	-0.49	0.39	5.7	5.67	1	-1.23	0.53	0.23	0.09
2	m3	6	1.52	0.4	5	-0.22	0.36	0.15	0.22	4	-1.67	0.53	0.16	0.07
3	m1	9	1	0.34	8	-0.65	0.44	0.64	0.35	7	-1.45	0.52	0.19	0.08
4	m13	12	0.72	0.25	11	-0.15	0.34	0.2	0.45	10	-1.4	0.55	0.2	0.09
5	m18	14	-0.1	0.16	13	0.06	0.17	0.55	1.87					
6	m9	16	0.78	0.19	15	-0.23	0.18	0.29	0.23					
7	m4	18	2.05	0.38	17	1.84	0.33	-0.89	0.13					
8	m16	21	1.38	0.36	20	0.4	0.31	-0.29	0.25	19	-1.52	0.54	0.18	0.08
9	m14	24	0.5	0.24	23	1.61	0.3	-3.21	1.49	22	-1.39	0.56	0.2	0.09
10	m6	27	0.52	0.23	26	-0.46	0.38	0.89	0.68	25	-1.39	0.55	0.2	0.09
11	m11	30	0.22	0.14	29	0.95	0.27	-4.27	2.85	28	-1.37	0.56	0.2	0.09
12	m7	33	1.67	0.42	32	1.61	0.35	-0.97	0.26	31	-1.38	0.54	0.2	0.09
13	m5	36	0.72	0.3	35	-0.91	0.48	1.27	0.54	34	-1.49	0.55	0.18	0.08
14	m2	39	0.2	0.13	38	-0.78	0.45	3.92	3.16	37	-1.27	0.53	0.22	0.09
15	m10	42	0.28	0.17	41	-0.88	0.46	3.15	2.22	40	-1.36	0.54	0.2	0.09
16	m15	45	0.46	0.21	44	-0.4	0.37	0.87	0.77	43	-1.37	0.55	0.2	0.09
17	m8	48	1.08	0.3	47	1.12	0.3	-1.04	0.37	46	-1.35	0.55	0.21	0.09
18	m17	51	0.92	0.27	50	0.18	0.32	-0.2	0.37	49	-1.42	0.55	0.19	0.09
19	v3a1	54	0.74	0.25	53	0.33	0.31	-0.44	0.46	52	-1.39	0.55	0.2	0.09
20	v3a3	57	1.02	0.39	56	-1.02	0.54	1	0.38	55	-1.47	0.51	0.19	0.08
21	v3a9	60	0.48	0.27	59	-1.09	0.53	2.29	1.21	58	-1.43	0.54	0.19	0.08
22	v3a17	63	1.31	0.35	62	1.09	0.32	-0.83	0.3	61	-1.39	0.54	0.2	0.09
23	v3b4	66	1.73	0.47	65	0.65	0.35	-0.38	0.24	64	-1.3	0.49	0.21	0.08
24	v3b12	69	0.2	0.13	68	1.23	0.28	-6.08	4.13	67	-1.38	0.56	0.2	0.09
25	v3b15	72	0.16	0.12	71	-1.75	0.96	10.74	9.23	70	-1.05	0.49	0.26	0.09
26	v3b16	75	1.13	0.4	74	-0.78	0.5	0.69	0.33	73	-1.41	0.5	0.2	0.08
27	v5a7	78	1.93	0.59	77	-0.43	0.48	0.22	0.21	76	-1.32	0.43	0.21	0.07
28	v5a10	81	0.7	0.27	80	-0.38	0.38	0.55	0.5	79	-1.46	0.56	0.19	0.09
29	v5a11	84	0.7	0.31	83	-0.8	0.5	1.14	0.59	82	-1.29	0.52	0.22	0.09
30	v5a14	87	0.45	0.22	86	1.09	0.29	-2.41	1.25	85	-1.39	0.56	0.2	0.09
31	v5a18	90	0.4	0.2	89	-0.35	0.36	0.87	0.89	88	-1.41	0.56	0.2	0.09
32	v5b2	93	0.63	0.25	92	-0.16	0.35	0.26	0.54	91	-1.42	0.56	0.19	0.09
33	v5b5	96	0.48	0.21	95	0.26	0.31	-0.53	0.69	94	-1.42	0.56	0.2	0.09
34	v5b6	99	0.85	0.28	98	0.01	0.34	-0.02	0.4	97	-1.47	0.56	0.19	0.09
35	v5b8	102	1.19	0.34	101	0.89	0.33	-0.75	0.33	100	-1.35	0.54	0.21	0.09
36	v5b13	105	1.84	0.5	104	1.91	0.43	-1.03	0.24	103	-1.46	0.55	0.19	0.08

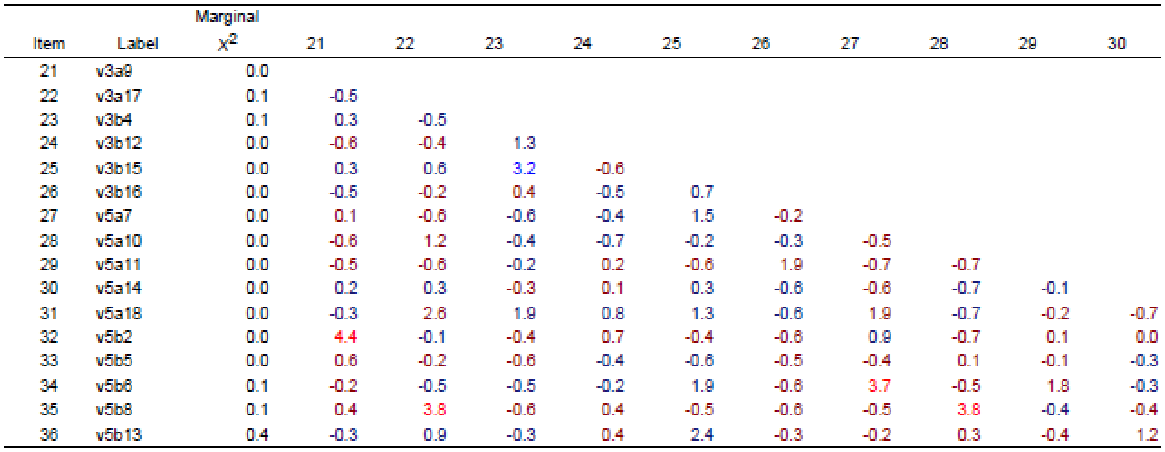
S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	Probability
1	m12	22.61	13	0.0465
2	m3	17.09	9	0.0472
3	m1	11.45	13	0.574
4	m13	14.86	13	0.3183
5	m18	23.88	17	0.1223
6	m9	11.44	13	0.575
7	m4	8.78	9	0.4593
8	m16	22.74	10	0.0117
9	m14	9.17	7	0.2403
10	m6	16.96	14	0.2578
11	m11	12.96	10	0.2252
12	m7	12.07	6	0.0604
13	m5	10.17	13	0.6808
14	m2	15.07	13	0.3022
15	m10	20.65	15	0.1479
16	m15	13.68	14	0.4751
17	m8	12.96	10	0.2254
18	m17	12.54	10	0.2497
19	v3a1	11.32	11	0.4183
20	v3a3	12.92	12	0.3767
21	v3a9	16.85	13	0.2057
22	v3a17	11.43	9	0.2467
23	v3b4	7.39	9	0.5978
24	v3b12	13.82	10	0.1809
25	v3b15	22.9	15	0.0861
26	v3b16	12.15	12	0.4349
27	v5a7	12.32	10	0.2634
28	v5a10	25.18	12	0.014
29	v5a11	11.31	14	0.6622
30	v5a14	8.37	10	0.5938
31	v5a18	11.57	15	0.7124
32	v5b2	11.26	14	0.6664
33	v5b5	13.07	13	0.4437
34	v5b6	18.8	12	0.0932
35	v5b8	10.87	9	0.2867
36	v5b13	10.22	7	0.176

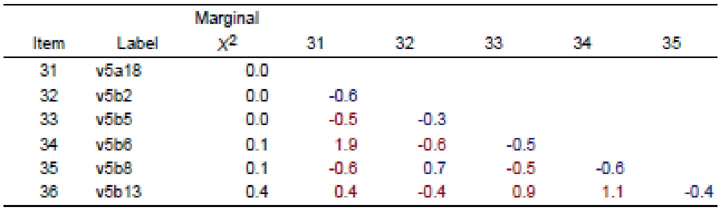
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1 (Back to TOC)

Item	Label	Marginal χ^2	1	2	3	4	5	6	7	8	9	10
1	m12	0.0										
2	m3	0.1	4.2									
3	m1	0.0	-0.1	0.5								
4	m13	0.0	4.7	0.3	-0.6							
5	m18	0.0	1.2	-0.6	-0.7	0.1						
6	m9	0.0	-0.4	-0.6	3.7	0.0	-0.1					
7	m4	0.1	2.7	1.5	-0.5	-0.6	-0.3	-0.1				
8	m16	0.0	4.0	-0.6	0.3	-0.7	-0.5	-0.7	-0.6			
9	m14	0.0	0.0	0.4	3.5	-0.6	1.9	-0.5	1.4	0.6		
10	m6	0.0	1.1	0.2	-0.3	-0.6	0.5	-0.7	1.3	3.7	-0.5	
11	m11	0.0	-0.3	0.1	2.4	0.2	-0.5	-0.6	-0.6	0.0	-0.5	-0.5
12	m7	0.0	1.9	0.8	-0.3	-0.7	-0.2	-0.6	-0.2	-0.6	-0.7	-0.6
13	m5	0.0	-0.6	-0.4	1.7	0.6	-0.6	-0.6	-0.3	-0.7	1.8	2.0
14	m2	0.0	-0.7	-0.2	0.7	-0.5	-0.1	-0.5	1.8	0.4	1.7	1.2
15	m10	0.0	0.5	-0.6	1.5	0.0	0.4	-0.1	0.2	0.2	2.0	1.3
16	m15	0.0	-0.7	-0.7	-0.7	-0.4	0.5	0.0	0.6	-0.5	0.4	-0.4
17	m8	0.0	0.7	-0.6	-0.3	-0.7	-0.2	-0.7	-0.3	-0.1	4.0	-0.4
18	m17	0.0	-0.7	1.2	-0.7	-0.2	-0.7	-0.7	-0.5	1.6	-0.6	-0.7
19	v3a1	0.0	1.2	0.2	0.2	-0.6	-0.7	-0.4	-0.1	-0.5	-0.7	-0.2
20	v3a3	0.0	5.8	3.6	-0.5	-0.5	-0.7	-0.3	0.7	-0.1	2.2	-0.6
21	v3a9	0.0	-0.3	-0.5	0.6	-0.6	-0.6	-0.4	0.4	-0.1	3.7	-0.6
22	v3a17	0.1	4.4	-0.4	-0.4	-0.5	-0.2	0.1	0.6	-0.5	0.9	-0.5
23	v3b4	0.1	4.4	-0.5	-0.6	1.8	0.9	-0.4	0.5	0.2	-0.6	-0.6
24	v3b12	0.0	5.0	-0.1	-0.6	1.1	-0.6	-0.2	0.9	-0.5	-0.1	0.8
25	v3b15	0.0	0.1	0.7	0.1	0.5	-0.6	0.7	0.9	-0.6	-0.2	-0.5
26	v3b16	0.0	4.2	-0.4	-0.4	-0.4	0.2	0.4	-0.1	2.3	0.6	-0.3
27	v5a7	0.0	1.8	2.0	-0.3	0.8	-0.5	-0.5	0.5	-0.2	-0.7	-0.5
28	v5a10	0.0	1.0	-0.4	0.2	-0.3	-0.5	0.5	0.2	0.5	-0.6	-0.7
29	v5a11	0.0	3.6	-0.2	-0.3	0.8	-0.4	-0.2	1.5	-0.3	-0.6	-0.4
30	v5a14	0.0	-0.2	1.6	-0.4	0.5	-0.6	0.0	0.1	-0.2	2.8	0.4
31	v5a18	0.0	0.0	-0.4	-0.4	0.6	17.4	-0.2	-0.0	0.5	-0.0	-0.7
32	v5b2	0.0	-0.1	-0.5	1.1	-0.2	-0.4	-0.2	0.1	-0.1	-0.1	-0.7
33	v5b5	0.0	3.1	-0.4	0.3	-0.1	0.1	3.4	1.5	0.9	1.3	-0.4
34	v5b6	0.1	3.3	-0.4	0.2	3.2	-0.4	-0.1	0.2	0.6	-0.6	2.9
35	v5b8	0.1	0.0	-0.4	-0.1	-0.2	-0.3	0.1	-0.0	0.8	0.1	-0.4
36	v5b13	0.4	2.0	0.3	-0.0	3.0	-0.0	0.3	0.0	0.0	-0.1	2.9

Item	Label	Marginal χ^2	11	12	13	14	15	16	17	18	19	20
11	m11	0.0										
12	m7	0.0	-0.7									
13	m5	0.0	-0.6	-0.0								
14	m2	0.0	0.1	0.7	-0.2							
15	m10	0.0	-0.7	-0.7	-0.6	-0.7						
16	m15	0.0	0.3	1.5	-0.3	0.4	-0.1					
17	m8	0.0	0.2	-0.2	-0.6	-0.3	-0.3	-0.3				
18	m17	0.0	-0.3	0.1	-0.3	0.7	-0.7	-0.6	-0.7			
19	v3a1	0.0	-0.3	-0.5	-0.1	0.0	1.9	0.1	0.3	-0.3		
20	v3a3	0.0	-0.5	-0.5	1.2	-0.5	-0.2	0.1	-0.1	-0.2	-0.2	
21	v3a9	0.0	0.1	-0.4	-0.1	0.1	-0.7	0.0	-0.5	-0.7	-0.5	-0.5
22	v3a17	0.1	0.4	1.9	0.6	1.8	-0.6	0.1	2.5	-0.5	-0.6	-0.3
23	v3b4	0.1	-0.2	-0.4	0.6	-0.6	-0.4	0.5	-0.7	-0.6	0.1	0.2
24	v3b12	0.0	-0.3	0.4	-0.1	-0.7	-0.7	-0.2	-0.7	1.2	-0.2	-0.5
25	v3b15	0.0	6.3	1.2	4.7	-0.5	-0.6	0.2	-0.7	-0.2	-0.3	0.3
26	v3b16	0.0	2.2	-0.4	0.4	-0.5	-0.6	1.4	-0.5	-0.6	0.0	3.9
27	v5a7	0.0	1.5	-0.0	-0.1	-0.6	-0.6	-0.2	-0.6	-0.1	2.2	0.1
28	v5a10	0.0	-0.2	-0.7	-0.2	0.4	4.6	0.9	-0.6	-0.1	-0.4	-0.3
29	v5a11	0.0	3.7	-0.5	-0.1	-0.6	0.5	-0.0	-0.3	-0.7	-0.3	0.7
30	v5a14	0.0	-0.1	0.1	1.2	-0.7	-0.7	0.1	-0.4	1.0	1.3	2.8
31	v5a18	0.0	-0.0	-0.6	0.3	-0.7	-0.5	0.2	0.1	0.1	-0.4	0.4
32	v5b2	0.0	-0.3	-0.0	-0.2	3.6	-0.6	-0.1	-0.7	-0.4	10.1	-0.2
33	v5b5	0.0	1.6	-0.5	6.6	-0.2	-0.5	-0.2	1.3	0.1	-0.5	-0.6
34	v5b6	0.1	0.1	-0.3	0.9	-0.6	0.7	1.2	-0.4	0.3	-0.4	0.8
35	v5b8	0.1	3.7	2.1	-0.0	-0.3	-0.3	-0.0	6.8	1.4	1.2	-0.3
36	v5b13	0.4	-0.1	-0.3	0.8	-0.3	0.5	0.4	1.7	0.9	0.5	-0.4



Item	Label	Marginal χ^2	21	22	23	24	25	26	27	28	29	30
21	v3a9	0.0										
22	v3a17	0.1	-0.5									
23	v3b4	0.1	0.3	-0.5								
24	v3b12	0.0	-0.6	-0.4	1.3							
25	v3b15	0.0	0.3	0.6	3.2	-0.6						
26	v3b16	0.0	-0.5	-0.2	0.4	-0.5	0.7					
27	v5a7	0.0	0.1	-0.6	-0.6	-0.4	1.5	-0.2				
28	v5a10	0.0	-0.6	1.2	-0.4	-0.7	-0.2	-0.3	-0.5			
29	v5a11	0.0	-0.5	-0.6	-0.2	0.2	-0.6	1.9	-0.7	-0.7		
30	v5a14	0.0	0.2	0.3	-0.3	0.1	0.3	-0.6	-0.6	-0.7	-0.1	
31	v5a18	0.0	-0.3	2.6	1.9	0.8	1.3	-0.6	1.9	-0.7	-0.2	-0.7
32	v5b2	0.0	4.4	-0.1	-0.4	0.7	-0.4	-0.6	0.9	-0.7	0.1	0.0
33	v5b5	0.0	0.6	-0.2	-0.6	-0.4	-0.6	-0.5	-0.4	0.1	-0.1	-0.3
34	v5b6	0.1	-0.2	-0.5	-0.5	-0.2	1.9	-0.6	3.7	-0.5	1.8	-0.3
35	v5b8	0.1	0.4	3.8	-0.6	0.4	-0.5	-0.6	-0.5	3.8	-0.4	-0.4
36	v5b13	0.4	-0.3	0.9	-0.3	0.4	2.4	-0.3	-0.2	0.3	-0.4	1.2



Item	Label	Marginal χ^2	31	32	33	34	35
31	v5a18	0.0					
32	v5b2	0.0	-0.6				
33	v5b5	0.0	-0.5	-0.3			
34	v5b6	0.1	1.9	-0.6	-0.5		
35	v5b8	0.1	-0.6	0.7	-0.5	-0.6	
36	v5b13	0.4	0.4	-0.4	0.9	1.1	-0.4

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)

-2loglikelihood:	5975.32
Akaike Information Criterion (AIC):	6185.32
Bayesian Information Criterion (BIC):	6499.31

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	36
Number of Dimensions	1

Appendix Z – jMetrik output for DIF Analysis

DIF by Sex/Gender

Item	Chi-square	p-value	Valid N	E.S.	95% CI	Class
m12	0.06	0.81	88	-0.23	(1.72, -2.18)	A
m3	0.01	0.9	85	0.14	(2.34, -2.05)	A
m1	0.13	0.72	97	-0.38	(1.66, -2.42)	A
m13	0.61	0.44	86	0.89	(3.05, -1.27)	A
m18	0.07	0.8	99	-0.26	(1.69, -2.22)	A
m9	0.83	0.36	89	0.9	(2.84, -1.04)	A
m4	0.4	0.53	67	-0.83	(1.71, -3.38)	A
m16	0.49	0.48	81	0.78	(3.03, -1.46)	A
m14	0.04	0.85	78	0.28	(3.05, -2.49)	A
m6	0	0.97	99	-0.04	(1.85, -1.93)	A
m11	8.71	0	90	4.41	(7.77, 1.04)	C+
m7	0.62	0.43	70	-1.15	(1.69, -3.99)	A
m5	3	0.08	90	-2.21	(0.23, -4.65)	A
m2	1.17	0.28	92	-1.09	(0.89, -3.08)	A
m10	3.67	0.06	94	-2.29	(0.00, -4.59)	A
m15	0.01	0.93	99	-0.08	(1.82, -1.98)	A
m8	0.04	0.85	75	0.26	(2.80, -2.28)	A
m17	0.31	0.58	90	0.58	(2.65, -1.49)	A
v3a1	0.84	0.36	75	-1.09	(1.25, -3.42)	A
v3a3	0	1	96	0.01	(2.00, -1.99)	A
v3a9	0.13	0.72	95	-0.36	(1.62, -2.34)	A
v3a17	1.24	0.27	64	-1.54	(1.07, -4.15)	A
v3b4	0	0.98	84	-0.03	(2.16, -2.22)	A
v3b12	0.35	0.56	84	0.66	(3.00, -1.67)	A
v3b15	0.24	0.63	99	0.44	(2.29, -1.40)	A
v3b16	0.05	0.82	96	0.25	(2.30, -1.79)	A
v5a7	0	0.95	76	-0.08	(2.20, -2.36)	A
v5a10	0.05	0.82	80	0.24	(2.34, -1.86)	A
v5a11	0.03	0.87	101	0.15	(2.03, -1.73)	A
v5a14	1.75	0.19	86	1.63	(4.02, -0.76)	A
v5a18	0.03	0.87	96	0.16	(2.00, -1.69)	A
v5b2	1.11	0.29	93	-1.2	(0.97, -3.38)	A
v5b5	2.67	0.1	88	-1.8	(0.29, -3.89)	A
v5b6	1.31	0.25	92	1.24	(3.43, -0.95)	A
v5b8	0	0.97	85	0.05	(2.36, -2.26)	A
v5b13	3.87	0.05	60	3.67	(7.75, -0.41)	B+

Matching Variable: Score
DIF Group Variable: Sex
Focal Group: Female
Reference Group: Male

DIF by Race/Ethnicity

Item	Chi-square	p-value	Valid N	E.S.	95% CI		Class
m12	0	0.99	56	-0.02	(2.79,	-2.84)	A
m3	0.22	0.64	58	0.7	(3.57,	-2.17)	A
m1	2.94	0.09	64	2.42	(5.32,	-0.49)	A
m13	0.79	0.37	61	1.14	(3.82,	-1.54)	A
m18	0.51	0.48	64	-1.01	(1.75,	-3.76)	A
m9	0.33	0.57	52	-0.9	(2.12,	-3.93)	A
m4	1.28	0.26	40	1.84	(5.12,	-1.44)	A
m16	1.63	0.2	52	-3.29	(1.35,	-7.93)	A
m14	3.4	0.07	44	3.16	(6.65,	-0.32)	A
m6	0.16	0.69	66	0.59	(3.27,	-2.1)	A
m11	0	0.95	54	0.09	(2.98,	-2.79)	A
m7	3.43	0.06	37	2.5	(5.67,	-0.66)	A
m5	0.07	0.8	60	0.33	(2.98,	-2.31)	A
m2	1.84	0.17	62	-1.95	(0.75,	-4.64)	A
m10	0.03	0.86	60	0.3	(3.44,	-2.85)	A
m15	8.76	0	58	-6.08	(-1.18,	-10.98)	C-
m8	0.78	0.38	43	-1.89	(2.22,	-6.01)	A
m17	8.97	0	41	6.69	(11.93,	1.44)	C+
v3a1	0.01	0.93	45	0.16	(3.54,	-3.22)	A
v3a3	1.51	0.22	63	-1.62	(1.08,	-4.31)	A
v3a9	0.01	0.91	55	-0.14	(2.51,	-2.8)	A
v3a17	2.28	0.13	41	2.85	(6.34,	-0.63)	A
v3b4	1.63	0.2	50	-1.6	(1.36,	-4.55)	A
v3b12	0.1	0.75	44	-0.75	(3.38,	-4.88)	A
v3b15	3.58	0.06	64	-2.44	(0.32,	-5.21)	A
v3b16	0.85	0.36	64	-1.42	(1.65,	-4.5)	A
v5a7	0.46	0.5	49	1.1	(4.33,	-2.14)	A
v5a10	2.66	0.1	53	-2.68	(0.63,	-5.98)	A
v5a11	0.03	0.86	66	0.2	(2.57,	-2.17)	A
v5a14	1.72	0.19	36	2.06	(5.52,	-1.41)	A
v5a18	1.29	0.26	66	1.4	(3.90,	-1.1)	A
v5b2	0.01	0.92	58	0.15	(3.15,	-2.85)	A
v5b5	0.77	0.38	56	1.35	(4.24,	-1.55)	A
v5b6	0.6	0.44	56	-1.25	(1.88,	-4.38)	A
v5b8	0.97	0.32	42	-1.95	(2.16,	-6.06)	A

Matching Variable: Score
DIF Group Variable: Race
Focal Group: Asian
Reference Group: White

Item	Chi-square	p-value	Valid N	E.S.	95% CI		Class
m12	1	0.32	8	Infinity	(NaN,	NaN)	A
m3	0.33	0.56	8	Infinity	(NaN,	NaN)	A
m1	3	0.08	8	Infinity	(NaN,	NaN)	A
m13	3	0.08	8	Infinity	(NaN,	NaN)	A
m18	0.6	0.44	8	Infinity	(NaN,	NaN)	A
m9	3	0.08	8	Infinity	(NaN,	NaN)	A
m4	NaN	NaN	0	NaN	(NaN,	NaN)	B-
m16	0.14	0.71	8	Infinity	(NaN,	NaN)	A
m14	NaN	NaN	0	NaN	(NaN,	NaN)	B-
m6	0.6	0.44	8	Infinity	(NaN,	NaN)	A
m11	0.6	0.44	8	Infinity	(NaN,	NaN)	A
m7	7	0.01	8	Infinity	(NaN,	NaN)	A
m5	1	0.32	8	Infinity	(NaN,	NaN)	A
m2	0.33	0.56	8	Infinity	(NaN,	NaN)	A
m10	1.67	0.2	8	Infinity	(NaN,	NaN)	A
m15	1	0.32	8	Infinity	(NaN,	NaN)	A
m8	0.14	0.71	8	Infinity	(NaN,	NaN)	A
m17	0.14	0.71	8	Infinity	(NaN,	NaN)	A
v3a1	0.14	0.71	8	Infinity	(NaN,	NaN)	A
v3a3	1.67	0.2	8	Infinity	(NaN,	NaN)	A
v3a9	1	0.32	8	Infinity	(NaN,	NaN)	A
v3a17	NaN	NaN	0	NaN	(NaN,	NaN)	B-
v3b4	0.33	0.56	8	Infinity	(NaN,	NaN)	A
v3b12	7	0.01	8	Infinity	(NaN,	NaN)	A
v3b15	3	0.08	8	Infinity	(NaN,	NaN)	A
v3b16	0.33	0.56	8	Infinity	(NaN,	NaN)	A
v5a7	1	0.32	8	Infinity	(NaN,	NaN)	A
v5a10	0.6	0.44	8	Infinity	(NaN,	NaN)	A
v5a11	1	0.32	8	Infinity	(NaN,	NaN)	A
v5a14	NaN	NaN	0	NaN	(NaN,	NaN)	B-
v5a18	3	0.08	8	Infinity	(NaN,	NaN)	A
v5b2	1.67	0.2	8	Infinity	(NaN,	NaN)	A
v5b5	0.14	0.71	8	Infinity	(NaN,	NaN)	A
v5b6	0.6	0.44	8	Infinity	(NaN,	NaN)	A

Matching Variable: Score
DIF Group Variable: Race
Focal Group: Black
Reference Group: White

Item	Chi-square	p-value	Valid N	E.S.	95% CI	Class
m12	0.46	0.5	22	-2.3	(3.61, -8.22)	A
m3	1.35	0.24	20	-3.18	(2.57, -8.94)	A
m1	2.84	0.09	22	-4.11	(1.09, -9.31)	A
m13	0.18	0.67	20	-2.38	(6.72, -11.48)	A
m18	0.02	0.9	22	-0.38	(4.82, -5.58)	A
m9	0.62	0.43	22	-2.2	(2.86, -7.25)	A
m4	0.03	0.87	16	-0.43	(4.42, -5.28)	A
m16	3.18	0.07	20	Infinity	(NaN, NaN)	A
m14	1.64	0.2	14	4.57	(11.02, -1.87)	A
m6	0.1	0.75	22	0.68	(5.01, -3.66)	A
m11	0	0.96	20	-0.14	(5.44, -5.73)	A
m7	0.38	0.54	8	2.15	(8.38, -4.08)	A
m5	1.36	0.24	22	-3.26	(2.10, -8.61)	A
m2	1.11	0.29	22	3.4	(9.36, -2.56)	A
m10	0.86	0.35	20	-2.86	(2.86, -8.58)	A
m15	1.36	0.24	26	-2.15	(1.87, -6.18)	A
m8	0	0.96	20	0.2	(6.81, -6.40)	A
m17	1.13	0.29	22	2.27	(7.04, -2.50)	A
v3a1	0.94	0.33	22	-3.18	(2.82, -9.19)	A
v3a3	1.54	0.21	22	-3.26	(1.51, -8.03)	A
v3a9	0.43	0.51	24	-1.45	(2.98, -5.87)	A
v3a17	0.05	0.82	12	-0.95	(6.37, -8.28)	A
v3b4	0.83	0.36	14	-Infinity	(NaN, NaN)	A
v3b12	0.3	0.58	14	1.1	(5.55, -3.34)	A
v3b15	4.91	0.03	24	Infinity	(NaN, NaN)	B+
v3b16	0.13	0.72	20	0.85	(6.13, -4.42)	A
v5a7	1.27	0.26	20	-3.33	(2.47, -9.13)	A
v5a10	0.1	0.75	20	-0.81	(4.10, -5.72)	A
v5a11	0.34	0.56	22	1.27	(5.48, -2.94)	A
v5a14	6.92	0.01	12	Infinity	(NaN, NaN)	B+
v5a18	0.74	0.39	24	-1.88	(2.72, -6.47)	A
v5b2	1.53	0.22	20	3.93	(10.05, -2.19)	A
v5b5	0.07	0.79	24	-0.6	(3.76, -4.96)	A
v5b6	0.33	0.56	16	-1.63	(4.34, -7.59)	A
v5b8	0.76	0.38	22	2.51	(7.63, -2.62)	A
v5b13	2.25	0.13	14	5.16	(11.43, -1.10)	A

Matching Variable: Score
DIF Group Variable: Race
Focal Group: Hispanic/Latino
Reference Group: White

Item	Chi-square	p-value	Valid N	E.S.	95% CI		Class
m12	2.64	0.1	47	2.43	(5.69,	-0.83)	A
m3	3.28	0.07	45	-3.16	(0.42,	-6.74)	A
m1	0.16	0.69	42	-0.59	(2.46,	-3.64)	A
m13	0.35	0.56	45	1.17	(4.89,	-2.55)	A
m18	0.1	0.75	47	-0.53	(2.65,	-3.72)	A
m9	0.95	0.33	45	-1.61	(1.61,	-4.82)	A
m4	0.75	0.39	38	-1.81	(2.25,	-5.88)	A
m16	0.08	0.78	47	-0.44	(2.67,	-3.54)	A
m14	1.98	0.16	31	-Infinity	(Nan,	NaN)	A
m6	0.07	0.79	47	0.48	3.76,	-2.79)	A
m11	0.77	0.38	40	-1.69	2.30,	-5.68)	A
m7	0.3	0.58	23	-1.15	3.08,	-5.38)	A
m5	0.36	0.55	40	1.33	5.36,	-2.71)	A
m2	0.05	0.83	45	0.37	3.64,	-2.91)	A
m10	0.07	0.79	45	0.45	3.92,	-3.01)	A
m15	0.01	0.93	47	-0.15	3.00,	-3.30)	A
m8	0.09	0.76	47	-0.51	2.73,	-3.76)	A
m17	0.01	0.9	39	0.16	3.02,	-2.69)	A
v3a1	0.06	0.81	47	0.39	3.54,	-2.77)	A
v3a3	0.4	0.53	45	-1.11	2.25,	-4.48)	A
v3a9	6.68	0.01	40	5.28	10.08,	0.48)	B+
v3a17	0.68	0.41	36	-1.86	2.46,	-6.17)	A
v3b4	4.94	0.03	39	-3.81	-0.10,	-7.52)	B-
v3b12	0.1	0.75	22	-1.14	5.23,	-7.51)	A
v3b15	2.4	0.12	45	3.75	8.37,	-0.88)	A
v3b16	0.18	0.67	45	-0.64	2.46,	-3.75)	A
v5a7	0.16	0.69	45	0.76	4.30,	-2.78)	A
v5a10	0.33	0.57	45	-0.96	2.29,	-4.21)	A
v5a11	0.02	0.88	47	0.25	3.50,	-2.99)	A
v5a14	0.15	0.7	40	-0.56	2.88,	-3.99)	A
v5a18	0.17	0.68	45	0.84	4.54,	-2.87)	A
v5b2	0.77	0.38	47	1.37	4.45,	-1.72)	A
v5b5	2.53	0.11	45	2.85	6.40,	-0.71)	A
v5b6	0.7	0.4	45	1.62	5.52,	-2.27)	A
v5b8	0.15	0.7	45	-0.54	2.60,	-3.68)	A
v5b13	0.19	0.66	29	1.15	5.91,	-3.60)	A

Matching Variable: Score
DIF Group Variable: Race
Focal Group: Other
Reference Group: White

DIF by Status

Item	Chi-square	p-value	Valid N	E.S.	(95% CI)		Class
m12	1.18	0.28	27	2.01	(5.82,	-1.81)	A
m3	0.01	0.91	23	-0.27	(4.30,	-4.84)	A
m1	1.01	0.32	31	1.96	(5.85,	-1.94)	A
m13	0.31	0.58	23	-1.6	(3.52,	-6.71)	A
m18	2.31	0.13	27	4.87	(11.10,	-1.37)	A
m9	0.28	0.6	27	-1.18	(2.92,	-5.27)	A
m4	0.42	0.51	27	-1.82	(3.13,	-6.78)	A
m16	5.18	0.02	27	-Infinity	(NaN,	NaN)	B-
m14	0.85	0.36	11	-Infinity	(NaN,	NaN)	A
m6	0.92	0.34	31	-2.12	(1.89,	-6.12)	A
m11	0.19	0.66	27	-0.75	(3.08,	-4.57)	A
m7	1.3	0.25	18	-4.98	(3.21,	-13.18)	A
m5	2.79	0.1	19	Infinity	(NaN,	NaN)	A
m2	0.36	0.55	31	1.08	4.62,	-2.46)	A
m10	0.3	0.58	27	1.2	5.52,	-3.12)	A
m15	0.02	0.88	22	-0.43	4.38,	-5.24)	A
m8	0.01	0.93	22	0.24	5.26,	-4.79)	A
m17	0.59	0.44	26	-1.58	2.33,	-5.49)	A
v3a1	0.04	0.83	27	-0.68	4.73,	-6.08)	A
v3a3	0.03	0.87	25	-0.36	3.67,	-4.40)	A
v3a9	2.64	0.1	27	3.65	8.65,	-1.34)	A
v3a17	0.33	0.56	23	-1.78	3.94,	-7.51)	A
v3b4	0.92	0.34	22	-2.58	2.40,	-7.57)	A
v3b12	0	0.98	15	0.07	6.04,	-5.91)	A
v3b15	0.1	0.75	31	0.65	4.52,	-3.22)	A
v3b16	3.11	0.08	31	-4.05	0.42,	-8.52)	A
v5a7	0.97	0.32	27	2.61	8.04,	-2.82)	A
v5a10	0.25	0.62	23	-1.17	3.34,	-5.69)	A
v5a11	0.06	0.81	27	-0.5	3.46,	-4.46)	A
v5a14	0	0.95	27	-0.18	4.75,	-5.10)	A
v5a18	0.56	0.45	31	1.43	5.18,	-2.31)	A
v5b2	0.38	0.54	27	1.25	5.31,	-2.82)	A
v5b5	0.32	0.57	27	1.08	4.85,	-2.69)	A
v5b6	0	0.95	31	-0.14	4.22,	-4.50)	A
v5b8	0.05	0.82	27	0.52	4.68,	-3.65)	A
v5b13	0.75	0.39	19	2.12	7.13,	-2.89)	A

Matching Variable: Score

DIF Group Variable: Status

Focal Group: 1st yr PhD

Reference Group: 5th yr PhD

Item	Chi-square	p-value	Valid N	E.S.	95% CI		Class
m12	0.01	0.91	12	0.31	(5.68,	-5.06)	A
m3	0.04	0.83	18	0.44	(4.50,	-3.61)	A
m1	0.33	0.57	16	1.74	(7.75,	-4.26)	A
m13	2.1	0.15	16	-3.37	(1.60,	-8.34)	A
m18	4.37	0.04	16	5.41	(10.89,	-0.07)	B+
m9	0.82	0.36	18	1.88	(6.03,	-2.28)	A
m4	0.67	0.41	5	-Infinity	NaN,	NaN)	A
m16	0.07	0.8	10	-0.68	(4.24,	-5.59)	A
m14	3.09	0.08	9	-Infinity	(NaN,	NaN)	A
m6	0.3	0.59	16	-1.1	(3.08,	-5.29)	A
m11	0.78	0.38	20	-2.06	(2.43,	-6.54)	A
m7	1.78	0.18	5	-Infinity	(NaN,	NaN)	A
m5	2.77	0.1	7	Infinity	(NaN,	NaN)	A
m2	0.01	0.93	18	-0.16	(3.66,	-3.98)	A
m10	1.11	0.29	18	2.21	(6.45,	-2.02)	A
m15	1.53	0.22	18	2.69	(7.20,	-1.81)	A
m8	0.02	0.89	7	0.43	(6.72,	-5.86)	A
m17	0.01	0.91	12	-0.31	(5.06,	-5.68)	A
v3a1	0.58	0.45	12	-2.38	(3.43,	-8.18)	A
v3a3	0.01	0.93	18	0.17	(4.13,	-3.78)	A
v3a9	0.23	0.63	18	-0.92	(2.92,	-4.76)	A
v3a17	0.62	0.43	11	-3.01	(3.65,	-9.67)	A
v3b4	0.17	0.68	18	1.05	(5.56,	-3.46)	A
v3b12	0.29	0.59	9	2.06	(9.32,	-5.20)	A
v3b15	0.36	0.55	18	1.25	(5.39,	-2.89)	A
v3b16	1.52	0.22	13	-3.01	(2.42,	-8.44)	A
v5a7	1.03	0.31	18	-1.93	(2.13,	-6.00)	A
v5a10	0.21	0.65	20	0.84	(4.45,	-2.78)	A
v5a11	0.49	0.49	16	-1.91	(2.98,	-6.79)	A
v5a14	0.58	0.45	12	-2.38	(3.43,	-8.18)	A
v5a18	0.36	0.55	18	-1.35	(2.98,	-5.69)	A
v5b2	0.64	0.42	20	1.69	(5.85,	-2.46)	A
v5b5	0.33	0.57	16	-1.45	(3.50,	-6.41)	A
v5b6	0	0.96	18	0.1	(4.25,	-4.06)	A
v5b8	3.15	0.08	14	Infinity	(NaN,	NaN)	A
v5b13	0.02	0.89	7	0.52	(7.28,	-6.23)	A

Matching Variable: Score
DIF Group Variable: Status
Focal Group: 2nd yr PhD
Reference Group: 5th yr PhD

Item	Chi-square	p-value	Valid N	E.S.	95% CI	Class
m12	0.24	0.62	12	1.45	(6.59, -3.68)	A
m3	0.36	0.55	10	1.32	(6.39, -3.76)	A
m1	0.46	0.5	13	-1.97	(3.14, -7.08)	A
m13	4.72	0.03	15	-Infinity	(NaN, NaN)	B-
m18	0.73	0.39	18	1.98	(6.38, -2.43)	A
m9	0.11	0.75	18	-0.83	(3.71, -5.36)	A
m4	0.1	0.75	11	-1.2	(5.26, -7.66)	A
m16	3.24	0.07	8	-Infinity	(NaN, NaN)	A
m14	2	0.16	3	-Infinity	(NaN, NaN)	A
m6	0.05	0.83	18	0.58	(5.36, -4.20)	A
m11	0.2	0.65	15	-1.18	(3.81, -6.16)	A
m7	0.5	0.48	3	-Infinity	(NaN, NaN)	A
m5	2.28	0.13	20	2.83	(6.85, -1.20)	A
m2	0.11	0.75	18	0.83	(5.36, -3.71)	A
m10	1.46	0.23	18	-2.62	(1.72, -6.96)	A
m15	0.64	0.42	17	1.51	(5.52, -2.50)	A
m8	0.06	0.81	7	-0.95	(6.27, -8.17)	A
m17	3.24	0.07	10	-Infinity	(NaN, NaN)	A
v3a1	0.06	0.81	18	-0.56	(3.77, -4.90)	A
v3a3	1.41	0.24	20	2.73	(7.12, -1.66)	A
v3a9	4.72	0.03	15	Infinity	(NaN, NaN)	B+
v3a17	0.36	0.55	10	-2.15	(4.45, -8.76)	A
v3b4	0	1	6	0	(7.98, -7.98)	A
v3b12	0.26	0.61	12	-1.47	(4.24, -7.17)	A
v3b15	0.8	0.37	11	-2.88	(3.37, -9.13)	A
v3b16	0.26	0.61	20	-1.05	(3.02, -5.11)	A
v5a7	1	0.32	10	2.94	(8.93, -3.04)	A
v5a10	1.06	0.3	12	-3.26	(2.75, -9.26)	A
v5a11	0.35	0.55	15	-1.66	(3.46, -6.78)	A
v5a14	0.14	0.71	17	0.69	(4.73, -3.34)	A
v5a18	0.98	0.32	17	2.02	(6.10, -2.06)	A
v5b2	0.28	0.6	18	1.22	(5.67, -3.24)	A
v5b5	0.31	0.58	14	1.32	(6.07, -3.42)	A
v5b6	3.21	0.07	10	Infinity	(NaN, NaN)	A
v5b8	0.07	0.79	14	-0.65	(4.02, -5.33)	A
v5b13	0.04	0.84	8	-0.68	(5.56, -6.91)	A

Matching Variable: Score
DIF Group Variable: Status
Focal Group: 3rd yr PhD
Reference Group: 5th yr PhD

Item	Chi-square	p-value	Valid N	E.S.	95% CI	Class
m12	0.63	0.43	13	-2.52	(3.14, -8.17)	A
m3	0.35	0.55	11	1.38	(6.11, -3.35)	A
m1	0	0.97	10	-0.1	(5.00, -5.19)	A
m13	1.47	0.23	5	Infinity	(NaN, NaN)	A
m18	1.19	0.28	12	2.78	(8.27, -2.71)	A
m9	0.04	0.85	17	0.43	(4.50, -3.64)	A
m4	1	0.32	2	-Infinity	(NaN, NaN)	A
m16	1.19	0.28	12	-2.33	(2.43, -7.10)	A
m14	1	0.32	2	-Infinity	(NaN, NaN)	A
m6	1.65	0.2	7	-Infinity	(NaN, NaN)	A
m11	2.86	0.09	11	-Infinity	(NaN, NaN)	A
m7	1	0.32	2	-Infinity	(NaN, NaN)	A
m5	2.52	0.11	17	3.17	(7.73, -1.39)	A
m2	0.06	0.81	13	0.52	(4.98, -3.94)	A
m10	0	0.95	15	0.13	(4.35, -4.09)	A
m15	0.38	0.54	13	1.46	(5.97, -3.05)	A
m8	1.99	0.16	12	-3.62	(1.40, -8.64)	A
m17	0.02	0.9	12	-0.32	(4.38, -5.03)	A
v3a1	0	0.95	12	-0.18	(4.77, -5.14)	A
v3a3	0.09	0.77	13	0.81	(5.72, -4.10)	A
v3a9	0	0.98	17	0.05	(3.63, -3.54)	A
v3a17	1	0.32	2	-Infinity	(NaN, NaN)	A
v3b4	0.15	0.69	8	1.2	(7.38, -4.98)	A
v3b12	0.5	0.48	3	Infinity	(NaN, NaN)	A
v3b15	1.47	0.23	5	-Infinity	(NaN, NaN)	A
v3b16	0.14	0.71	15	-0.69	(3.32, -4.71)	A
v5a7	0.01	0.94	11	-0.25	(5.57, -6.06)	A
v5a10	0.01	0.92	8	0.43	(8.09, -7.23)	A
v5a11	1.77	0.18	13	4.84	(11.42, -1.75)	A
v5a14	0.15	0.69	7	1.2	(7.12, -4.72)	A
v5a18	0.36	0.55	8	-2.15	(4.23, -8.54)	A
v5b2	0.09	0.77	13	-0.72	(4.02, -5.46)	A
v5b5	0.26	0.61	10	1.72	(7.69, -4.24)	A
v5b6	0.83	0.36	10	2.52	(7.87, -2.84)	A
v5b8	0	0.97	10	0.11	(5.56, -5.35)	A
v5b13	1	0.32	2	Infinity	(NaN, NaN)	A

Matching Variable: Score
DIF Group Variable: Status
Focal Group: 4th yr PhD
Reference Group: 5th yr PhD

Appendix AA – IRTPRO output for Dimensionality (Vignettes only)

Item Analysis – Combined Sample

Item	Option	Difficulty	Std. Dev.	Discrimin.
v3a1	Overall	0.6054	0.4904	0.5796
v3a3	Overall	0.3946	0.4904	0.3289
v3a9	Overall	0.3673	0.4837	0.3597
v3a17	Overall	0.6871	0.4653	0.6105
v3b4	Overall	0.6054	0.4904	0.5796
v3b12	Overall	0.7279	0.4466	0.5863
v3b15	Overall	0.3333	0.473	0.0913
v3b16	Overall	0.4218	0.4955	0.5755
v5a7	Overall	0.4694	0.5008	0.6036
v5a10	Overall	0.449	0.4991	0.5131
v5a11	Overall	0.4082	0.4932	0.5317
v5a14	Overall	0.6803	0.468	0.605
v5a18	Overall	0.4558	0.4997	0.4209
v5b2	Overall	0.4694	0.5008	0.5561
v5b5	Overall	0.5306	0.5008	0.5376
v5b6	Overall	0.4898	0.5016	0.6141
v5b8	Overall	0.6054	0.4904	0.7147
v5b13	Overall	0.6667	0.473	0.8395

```
=================================================================
  TEST LEVEL STATISTICS – Combined Sample
=================================================================
Number of Items = 18
Number of Examinees = 147
Min = 0.0000
Max = 17.0000
Mean = 9.3673
Median = 10.0000
Standard Deviation = 4.4279
Interquartile Range = 5.0000
Skewness = -0.6211
Kurtosis = -0.2274
KR21 = 0.8162
=================================================================
```

```
                    RELIABILITY ANALYSIS – Combined Sample
==================================================================
```

Method	Estimate	95% Conf. Int.		SEM
Guttman's L2	0.8361	(0.7949,	0.8723)	1.7988
Coefficient Alpha	0.8298	(0.7870,	0.8674)	1.833
Feldt-Gilmer	0.8339	(0.7921,	0.8705)	1.811
Feldt-Brennan	0.8322	(0.7900,	0.8692)	1.8201
Raju's Beta	0.8298	(0.7870,	0.8674)	1.833

RELIABILITY IF ITEM DELETED – Combined Sample

Item	L2	Alpha	F-G	F-B	Raju
v3a1	0.8263	0.8195	0.8242	0.8222	0.8195
v3a3	0.8357	0.83	0.8338	0.8322	0.83
v3a9	0.835	0.8287	0.8327	0.831	0.8287
v3a17	0.8261	0.8192	0.8239	0.8218	0.8192
v3b4	0.8263	0.8195	0.8239	0.822	0.8195
v3b12	0.8277	0.8207	0.8254	0.8234	0.8207
v3b15	0.8431	0.8389	0.8409	0.84	0.8389
v3b16	0.8264	0.8195	0.8242	0.8222	0.8195
v5a7	0.825	0.8181	0.8228	0.8207	0.8181
v5a10	0.8289	0.8221	0.8265	0.8246	0.8221
v5a11	0.8287	0.8215	0.8264	0.8243	0.8215
v5a14	0.8259	0.8193	0.8237	0.8218	0.8193
v5a18	0.8324	0.8261	0.8301	0.8284	0.8261
v5b2	0.8271	0.8202	0.8251	0.823	0.8202
v5b5	0.8281	0.821	0.8256	0.8237	0.821
v5b6	0.8245	0.8176	0.8221	0.8201	0.8176
v5b8	0.8207	0.8137	0.8182	0.8163	0.8137
v5b13	0.8161	0.8095	0.8135	0.8117	0.8095

```
=============================================================================
L2: Guttman's lambda-2
Alpha: Coefficient alpha
F-G: Feldt-Gilmer coeff.
F-B: Feldt-Brennan coef.
Raju: Raju's beta coeff.
```

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - Vig Items; EFA(1)

Item	*Label*		*a*	*s.e.*		*c*	*s.e*	*b*	*s.e.*
1	v3a1	2	0.68	0.27	1	0.68	0.2	-0.99	0.42
2	v3a3	4	0.83	0.28	3	-0.36	0.19	0.43	0.25
3	v3a9	6	0.43	0.23	5	-0.43	0.18	1.01	0.65
4	v3a17	8	1.16	0.38	7	1.32	0.27	-1.13	0.31
5	v3b4	10	1.07	0.33	9	0.89	0.23	-0.83	0.27
6	v3b12	12	0.15	0.27	11	1.5	0.23	-9.74	17.22
7	v3b15	14	-0.53	0.24	13	-0.5	0.19	-0.96	0.54
8	v3b16	16	1.13	0.35	15	-0.16	0.2	0.14	0.18
9	v5a7	18	1.39	0.37	17	0.19	0.22	-0.14	0.16
10	v5a10	20	0.68	0.26	19	0.06	0.19	-0.09	0.28
11	v5a11	22	0.68	0.26	21	-0.15	0.19	0.22	0.29
12	v5a14	24	0.4	0.28	23	1.34	0.23	-3.38	2.31
13	v5a18	26	0.5	0.24	25	0.1	0.18	-0.19	0.38
14	v5b2	28	0.48	0.23	27	0.24	0.19	-0.51	0.46
15	v5b5	30	0.56	0.25	29	0.6	0.2	-1.07	0.55
16	v5b6	32	1.11	0.34	31	0.41	0.22	-0.37	0.21
17	v5b8	34	0.96	0.34	33	1.11	0.24	-1.17	0.38
18	v5b13	36	1.89	0.62	35	2.07	0.46	-1.1	0.23

Factor Loadings for Group 1

Item	**Label**	**λ*1***	***s.e.***
1	v3a1	0.37	0.21
2	v3a3	0.44	0.2
3	v3a9	0.24	0.21
4	v3a17	0.56	0.21
5	v3b4	0.53	0.2
6	v3b12	0.09	0.27
7	v3b15	-0.3	0.21
8	v3b16	0.55	0.2
9	v5a7	0.63	0.17
10	v5a10	0.37	0.21
11	v5a11	0.37	0.2
12	v5a14	0.23	0.25
13	v5a18	0.28	0.21
14	v5b2	0.27	0.21
15	v5b5	0.31	0.22
16	v5b6	0.55	0.2
17	v5b8	0.49	0.22
18	v5b13	0.74	0.19

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on Monte Carlo estimated loglikelihood (95% CI)

-2loglikelihood:	2811.63 ± 0.57
Akaike Information Criterion (AIC):	2883.63 ± 0.57
Bayesian Information Criterion (BIC):	2991.29 ± 0.57

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	18
Number of Dimensions	1

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - Vig Items; EFA(2)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e*		*c*	*s.e.*
1	v3a1	2	2.19	0.7	3	-0.39	0.39	1	1.09	0.35
2	v3a3	5	0.26	0.25	6	0.88	0.31	4	-0.37	0.2
3	v3a9	8	0.51	0.25	9	0.14	0.24	7	-0.44	0.18
4	v3a17	11	0.5	0.31	12	1.14	0.37	10	1.36	0.28
5	v3b4	14	0.93	0.33	15	0.68	0.3	13	0.91	0.24
6	v3b12	17	0.47	0.32	18	-0.17	0.29	16	1.55	0.25
7	v3b15	20	-0.22	0.24	21	-0.53	0.26	19	-0.51	0.19
8	v3b16	23	0.72	0.28	24	0.84	0.3	22	-0.16	0.2
9	v5a7	26	0.48	0.32	27	1.69	0.55	25	0.24	0.25
10	v5a10	29	0.36	0.24	30	0.57	0.25	28	0.06	0.19
11	v5a11	32	0.55	0.27	33	0.41	0.25	31	-0.15	0.19
12	v5a14	35	0.69	0.32	36	0	0.29	34	1.39	0.25
13	v5a18	38	0.12	0.24	39	0.59	0.24	37	0.11	0.19
14	v5b2	41	2.51	1.42	42	-0.9	0.65	40	0.35	0.32
15	v5b5	44	0.25	0.26	45	0.5	0.27	43	0.61	0.2
16	v5b6	47	0.58	0.27	48	0.96	0.32	46	0.42	0.22
17	v5b8	50	0.35	0.3	51	0.96	0.35	49	1.16	0.26
18	v5b13	53	1.58	0.59	54	1.23	0.49	52	2.13	0.5

Factor Loadings for Group 1

Item	**Label**	***λ1***	*s.e.*	**λ2**	s.e.
1	v3a1	0.78	0.16	-0.14	0.22
2	v3a3	0.14	0.21	0.45	0.22
3	v3a9	0.29	0.22	0.08	0.23
4	v3a17	0.24	0.23	0.54	0.21
5	v3b4	0.45	0.21	0.33	0.21
6	v3b12	0.27	0.28	-0.1	0.28
7	v3b15	-0.12	0.22	-0.29	0.22
8	v3b16	0.35	0.2	0.42	0.2
9	v5a7	0.2	0.22	0.69	0.2
10	v5a10	0.19	0.22	0.31	0.21
11	v5a11	0.3	0.22	0.22	0.22
12	v5a14	0.38	0.26	0	0.27
13	v5a18	0.06	0.23	0.32	0.21
14	v5b2	0.79	0.22	-0.28	0.21
15	v5b5	0.14	0.24	0.28	0.23
16	v5b6	0.29	0.21	0.47	0.21
17	v5b8	0.18	0.25	0.49	0.23
18	v5b13	0.6	0.21	0.47	0.21

Likelihood-based Values and Goodness of Fit Statistics

Stats based on MC est. loglikelihood (95% CI)	
-2loglikelihood:	2771.72 ± 0.88
Akaike Information Criterion (AIC):	28879.72 ± 0.88
Bayesian Information Criterion (BIC):	3041.2 ± 0.88

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	18
Number of Dimensions	2

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - Vig Items; EFA(3)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e*		*a3*	*s.e.*		*c*	*s.e.*
1	v3a1	2	1.13	0.52	3	2.04	0.68	4	-0.61	0.46	1	1.14	0.4
2	v3a3	6	2.66	1.77	7	-0.86	0.85	8	1.04	0.81	5	-0.67	0.42
3	v3a9	10	0.28	0.24	11	0.39	0.24	12	0.14	0.24	9	-0.43	0.18
4	v3a17	14	0.36	0.33	15	0.49	0.32	16	1.15	0.38	13	1.4	0.29
5	v3b4	18	1.21	0.43	19	0.59	0.32	20	0.41	0.3	17	0.99	0.26
6	v3b12	22	-0.06	0.31	23	0.54	0.32	24	-0.13	0.3	21	1.57	0.25
7	v3b15	26	-0.27	0.25	27	-0.17	0.25	28	-0.45	0.26	25	-0.51	0.19
8	v3b16	30	1.32	0.42	31	0.28	0.29	32	0.63	0.31	29	-0.16	0.22
9	v5a7	34	0.58	0.34	35	0.37	0.35	36	1.68	0.55	33	0.26	0.25
10	v5a10	38	-0.07	0.27	39	0.61	0.3	40	0.76	0.31	37	0.08	0.2
11	v5a11	42	0.61	0.27	43	0.29	0.25	44	0.3	0.26	41	-0.15	0.19
12	v5a14	46	-0.18	0.34	47	1.05	0.39	48	0.07	0.32	45	1.54	0.29
13	v5a18	50	-0.29	0.29	51	0.38	0.3	52	0.8	0.32	49	0.13	0.2
14	v5b2	54	1.01	0.51	55	2.02	0.97	56	-0.96	0.6	53	0.32	0.33
15	v5b5	58	0.15	0.26	59	0.24	0.25	60	0.5	0.27	57	0.62	0.2
16	v5b6	62	0.61	0.32	63	0.39	0.29	64	0.89	0.32	61	0.44	0.22
17	v5b8	66	-0.06	0.34	67	0.61	0.35	68	1.16	0.39	65	1.28	0.3
18	v5b13	70	0.8	0.45	71	1.57	0.62	72	1.23	0.54	69	2.25	0.56

Factor Loadings for Group 1

Item	Label	λ*1*	*s.e.*	λ2	s.e.	λ3	s.e.
1	v3a1	0.38	0.22	0.69	0.17	-0.21	0.22
2	v3a3	0.77	0.22	-0.25	0.26	0.3	0.21
3	v3a9	0.16	0.23	0.22	0.22	0.08	0.23
4	v3a17	0.17	0.25	0.23	0.24	0.54	0.21
5	v3b4	0.55	0.23	0.27	0.22	0.19	0.23
6	v3b12	-0.03	0.29	0.3	0.28	-0.07	0.29
7	v3b15	-0.15	0.23	-0.09	0.23	-0.25	0.23
8	v3b16	0.59	0.2	0.12	0.21	0.28	0.21
9	v5a7	0.23	0.22	0.15	0.23	0.68	0.2
10	v5a10	-0.04	0.24	0.31	0.22	0.39	0.23
11	v5a11	0.33	0.22	0.16	0.23	0.16	0.23
12	v5a14	-0.09	0.28	0.52	0.24	0.04	0.27
13	v5a18	-0.15	0.24	0.19	0.24	0.41	0.22
14	v5b2	0.34	0.21	0.68	0.22	-0.32	0.22
15	v5b5	0.08	0.25	0.13	0.24	0.28	0.23
16	v5b6	0.3	0.24	0.19	0.23	0.43	0.21
17	v5b8	-0.03	0.26	0.29	0.25	0.54	0.21
18	v5b13	0.29	0.25	0.57	0.22	0.45	0.22

Likelihood-based Values and Goodness of Fit Statistics	
Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	2746.18 ± 1.20
Akaike Information Criterion (AIC):	2890.18 ± 1.20
Bayesian Information Criterion (BIC):	3105.49 ± 1.20

Summary of the Data and Control Parameters	
Sample Size	147
Number of Items	18
Number of Dimensions	3

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - Vig Items; CFA(2)

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*c*	*s.e.*
1	v3a1		0	-----	2	0.76	0.3	1	0.69	0.2
2	v3a3		0	-----	4	0.63	0.26	3	-0.34	0.18
3	v3a9		0	-----	6	0.46	0.24	5	-0.43	0.18
4	v3a17	8	2.28	1.15		0	-----	7	1.87	0.69
5	v3b4		0	-----	10	1.02	0.35	9	0.89	0.23
6	v3b12		0	-----	12	0.29	0.29	11	1.51	0.23
7	v3b15	14	-0.49	0.28		0	-----	13	-0.51	0.19
8	v3b16	16	0.62	0.3		0	-----	15	-0.12	0.19
9	v5a7		0	-----	18	1.16	0.38	17	0.19	0.21
10	v5a10		0	-----	20	0.69	0.26	19	0.07	0.19
11	v5a11		0	-----	22	0.62	0.25	21	-0.14	0.19
12	v5a14		0	-----	24	0.57	0.3	23	1.38	0.24
13	v5a18	26	0.96	0.44		0	-----	25	0.12	0.21
14	v5b2		0	-----	28	0.61	0.26	27	0.26	0.19
15	v5b5		0	-----	30	0.55	0.25	29	0.61	0.2
16	v5b6		0	-----	32	1.11	0.37	31	0.42	0.22
17	v5b8		0	-----	34	0.86	0.32	33	1.09	0.24
18	v5b13		0	-----	36	3.24	2.03	35	3.03	1.5

Factor Loadings for Group 1

Item	Label	λ*1*	*s.e.*	λ2	s.e.
1	v3a1	0	0	0.41	0.23
2	v3a3	0	0	0.35	0.22
3	v3a9	0	0	0.26	0.22
4	v3a17	0.8	0.25	0	0
5	v3b4	0	0	0.51	0.22
6	v3b12	0	0	0.17	0.28
7	v3b15	-0.28	0.25	0	0
8	v3b16	0.34	0.25	0	0
9	v5a7	0	0	0.56	0.21
10	v5a10	0	0	0.38	0.21
11	v5a11	0	0	0.34	0.21
12	v5a14	0	0	0.32	0.26
13	v5a18	0.49	0.29	0	0
14	v5b2	0	0	0.34	0.22
15	v5b5	0	0	0.31	0.22
16	v5b6	0	0	0.55	0.22
17	v5b8	0	0	0.45	0.23
18	v5b13	0	0	0.89	0.2

Likelihood-based Values and Goodness of Fit Statistics

Stats based on MC est. loglikelihood (95% CI)	
-2loglikelihood:	2837.61 ± 0.79
Akaike Information Criterion (AIC):	2909.61 ± 0.79
Bayesian Information Criterion (BIC):	3017.27 ± 0.79

Summary of the Data and Control Parameters

Sample Size	147
Number of Items	18
Number of Dimensions	2

IRTPRO Version 4.2
Output generated by IRTPRO estimation engine Version 5.20 (64-bit)
Project: Combined Sample - Vig Items; Bifactor

Item	*Label*		*a1*	*s.e.*		*a2*	*s.e.*		*a3*	*s.e.*		*c*	*s.e.*
1	v3a1	2	0.7	0.39	3	2.29	1.04		0	-----	1	1.14	0.42
2	v3a3	5	1.07	0.35	6	-0.28	0.29		0	-----	4	-0.38	0.21
3	v3a9	8	0.37	0.23	9	0.34	0.26		0	-----	7	-0.43	0.18
4	v3a17	11	1.77	0.83		0	-----	12	2.15	1.38	10	2.15	0.88
5	v3b4	14	1.07	0.34	15	0.51	0.29		0	-----	13	0.93	0.24
6	v3b12	17	0.05	0.28	18	0.5	0.33		0	-----	16	1.55	0.25
7	v3b15	20	-0.54	0.25		0	-----	21	-0.17	0.32	19	-0.51	0.19
8	v3b16	23	1.12	0.34		0	-----	24	0.03	0.43	22	-0.14	0.21
9	v5a7	26	1.74	0.52	27	-0.35	0.34		0	-----	25	0.26	0.25
10	v5a10	29	0.66	0.26	30	0.12	0.26		0	-----	28	0.07	0.19
11	v5a11	32	0.63	0.25	33	0.25	0.25		0	-----	31	-0.15	0.19
12	v5a14	35	0.28	0.29	36	0.69	0.32		0	-----	34	1.41	0.25
13	v5a18	38	0.51	0.27		0	-----	39	0.84	0.53	37	0.12	0.21
14	v5b2	41	0.37	0.37	42	2.3	0.83		0	-----	40	0.33	0.3
15	v5b5	44	0.54	0.26	45	0	0.26		0	-----	43	0.61	0.2
16	v5b6	47	1.15	0.34	48	0.07	0.28		0	-----	46	0.43	0.22
17	v5b8	50	0.96	0.32	51	-0.06	0.3		0	-----	49	1.14	0.25
18	v5b13	53	1.87	0.66	54	0.94	0.48		0	-----	52	2.2	0.54

Factor Loadings for Group 1

Item	Label	*λ1*	*s.e.*	λ2	s.e.	λ3	s.e.
1	v3a1	0.24	0.18	0.78	0.22	0	0
2	v3a3	0.53	0.21	-0.14	0.24	0	0
3	v3a9	0.21	0.21	0.19	0.24	0	0
4	v3a17	0.54	0.19	0	0	0.66	0.3
5	v3b4	0.52	0.2	0.25	0.22	0	0
6	v3b12	0.03	0.27	0.28	0.29	0	0
7	v3b15	-0.3	0.22	0	0	-0.09	0.3
8	v3b16	0.55	0.2	0	0	0.02	0.36
9	v5a7	0.71	0.18	-0.14	0.22	0	0
10	v5a10	0.36	0.21	0.07	0.24	0	0
11	v5a11	0.35	0.21	0.14	0.23	0	0
12	v5a14	0.15	0.26	0.37	0.25	0	0
13	v5a18	0.26	0.22	0	0	0.43	0.37
14	v5b2	0.13	0.21	0.8	0.18	0	0
15	v5b5	0.3	0.22	0	0.25	0	0
16	v5b6	0.56	0.19	0.04	0.23	0	0
17	v5b8	0.49	0.21	-0.03	0.26	0	0
18	v5b13	0.69	0.19	0.35	0.24	0	0

Likelihood-based Values and Goodness of Fit Statistics	
Statistics based on Monte Carlo estimated loglikelihood (95% CI)	
-2loglikelihood:	2770.90 ± 1.07
Akaike Information Criterion (AIC):	2878.90 ± 1.07
Bayesian Information Criterion (BIC):	3040.39 ± 1.07

Summary of the Data and Control Parameters	
Sample Size	147
Number of Items	18
Number of Dimensions	3

Appendix AB – Scored *p*-Value Definitions

ID	Score	*p*-value definition
1	0	cv
2	1	I don't know
3	0	
4	2	A measure of how likely a data set is to exhibit a correlation
5	0	cv
6	*	a numerical representation of statistical significance between two groups
7	1	I really have no idea...
8	3	The p-value is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event
9	1	No idea. I remember the term from Statistics in community college, though.
10	2	represents the probability that the null hypothesis is true
11	3	it's the probability of finding the observed results when the null hypothesis is true. It helps determine the significance of results.
12	*	probability
13	0	
14	2	determines the significance of data based on students t-test
15	3	The probability of obtaining a statistic as extreme or more extreme as the one you got
16	3	The probability of sampling a score at least as extreme as the observed sample
17	0	
18	3	A measure of the likelihood that a given phenomena would appear randomly
19	2	Some sort of statistical confidence value? Have never actually learned this! But I've seen it in more stats-heavy literature.
20	2	a probability against the null hypothesis
21	*	a measure of statistical significance
22	*	measure of significance
23	3	P-value refers to the confidence interval at which we can reject the null hypothesis
24	2	probability of an effect occurred by random chance
25	3	A measure of confidence in a statistic. Generally p values &lt;.05 are considered "good"and one can reject the null hypothesis
26	2	A p-value is the probability that an association is due to chance.
27	2	The p-value is the probability that you obtained a set of results by chance alone given that the null hypothesis is true.
28	*	a number that is generated to show the significance of a statistical finding
29	2	P-value is the probability of the event occurring due to chance.
30	*	A value of statistical significance
31	2	A p-value is the probability that the effect seen is due to random chance.

32	2	probability that a null hypothesis is true, often used to infer population statistics from a sample
33	3	An indication of significance, with a smaller number (i.e. less than 0.01) indicating a finding that is not likely to be due to chance alone.
34	2	The probability that the null hypothesis is false, or that your data is not statistically significantly different.
35	3	The probability of obtaining a test statistic as extreme or more extreme than the one obtained given ramdom chance
36	*	how significant your result is, the actuality that your results are not random
37	3	A p-value is a statistical construct that represents the statistical significance of a given calculation, or the likelihood that you would find the same result by random chance. The lower the p-value, the more likely that the result was not from random chance.
38	3	A p-value is a statistical representation of the likelihood of a series of results to relative to some null hypothesis. A low p-value indicates that the collected results are highly unlikely based on the null hypothesis.
39	*	It's a statistical measure of how confident you are in your results
40	2	A p-value is a way to describe the probability that a statistical result was observed by chance alone
41	2	A p-value indicates that the likelihood of an event occurring is p, a set value ranging between 0 and 1.
42	*	indicator of statistical significance for a given study
43	2	the probability of finding that result when the null hypothesis is true
44	2	The likelihood that the hypothesis occurred by chance
45	2	A statistic which indicates the probability that a relationship occurred due to chance.
46	0	cv
47	3	A value indicating the likelihood that a phenomenon is not just a result of chance.
48	4	A p-value is the probability that the statistical findings would be as or more extreme given the null hypothesis.
49	4	The probability of getting data at least this extreme, under the assumption that the null hypothesis is true. p(x|H0=TRUE)
50	*	A statistical measure that gives indication of whether of difference or relationships is statistically significant.
51	3	A p-value is the probability that the value you are computing from the data is greater/less/not equal (depending if it is one or two tailed) than its observed value, if the null hypothesis were correct.
52	2	A statistic representing the probability of a false positive for a given statistical hypothesis
53	2	A p-value is a value that indicates the likelihood that an observed effect (or a larger/more extreme effect than the one observed) occurred by chance.

54	*	Significance of finding
55	*	A p-value helps me determine the significance of a result
56	4	The probability that an event, or a more extreme event, occurred under some null hypothesis.
57	3	The probability that the observed F-ratio (or higher) given that the null hypothesis is true (or if true PRE=0).
58	4	It's the probability of observing your outcome or a more extreme one, given a null hypothesis.
59	2	tells you likelihood the results you're seeing are due to chance only
60	2	the probability that you would have retrieved a result by simple chance
61	3	A statistic indicating the likelihood in which a given result (i.e. mean) is equal to or below the "true" one (i.e. "true" mean), if samples were taken indefinitely.
62	2	A p-value is the result of an F test, an arbitrary alpha value indicating statistical significance; it means that the observation has a 5% chance that it was random if the p is set at .05.
63	2	The probability of a type I error (null hypothesis rejected when it was correct) in frequentist statistics.
64	2	The probability of falsely rejecting the null hypothesis (aka making a type 1 error).
65	2	A p-value is a number generated to indicate the likelihood that the conclusions drawn from the experiment are actually false and due to random chance.
66	2	A p-value is the probability that the results of an experiment could have happened by chance.
67	*	A calculated probability
68	*	a p-value determines the likelihood that a statistical test is significant or not.
69	3	A p-value is the probability that a given value as or more extreme would be observed if selected at random from the distribution of possible values
70	2	The probability that a null hypothesis will be accepted even if it is false.
71	2	A measure of the confidence that one might correctly reject a hull hypothesis when comparing two or more data sets.
72	*	a number that defines the threshold for statistical significance in many statistical tests.
73	2	the probability of rejecting the null-hypothesis
74	2	A p-value is a measure of statistical significance that represents the percentage chance of any result to have happened randomly.
75	*	p-value tells you the statistical significance of the results of hypothesis testing.
76	3	The probability that an observed measurement/data set fits a tested probability distribution by random chance rather than because of some underlying feature of the system.
77	2	Probability of observing said result if the null hypothesis is true
78	2	the probability that we observe such value, conditional on the null hypothesis is true.

79	4	The probability that a value is more extreme than the value observed in the data given your null hypothesis is "true"
80	2	is the probability that a type I error has occurred
81	2	A p-value tells you the probability that the result you are seeing was due to random chance.
82	2	the probability of a result being caused by random chance
83	2	the probability that an observed event is rare
84	2	the probability that a null-hypothesis is true
85	3	A number between 0 and 1 that represents the probability that a distribution is consistent with the Null Hypothesis. A p-value < 0.05 typically indicates a significant non-random correlation in the distribution.
86	3	A proportion of a distribution used to determine how likely a value (data point) is to belong to that distribution
87	3	A representation of the chance that the data can be explained by random variation of the data. A small p-value allows the null hypothesis to be discarded.
88	2	A p-value is a value that indicates the probability of a particular sampled observation.
89	2	Measure of the probability that the observed results were generated by chance
90	*	the level of statistical significance reported in an analysis. i.e. if there is a p-value of 0.05, there is a 5% chance that the results are due to chance.
91	4	p-value is the probability that a observed result is the same or greater given that the null hypothesis is true
92	2	The probability of getting the observed outcome when the null hypothesis is true.
93	2	Based on a normal distribution, a p-value represents the probability that the observed value varies from expected values (or the null hypothesis in hypothesis testing).
94	2	the p-value is the probability that a given result occurred by chance, rather than as a result of the instrument being studied.
95	4	The p-value is used In statistical hypothesis testing in order to quantify the idea of statistical significance of evidence. p-value is the probability under the context of a given statistical model that, when the null hypothesis is true, the statistical summary (such as the sample mean difference between two compared groups) would be the same as or of greater magnitude than the actual observed results.
96	*	A p-value is a number that calculated by some statistic methods, indicating how confident it is of us to say the two group of samples are related.
97	4	The probability that data would produce a statistic as or more extreme than observed, under a certain hypothesis.
98	2	The p-value indicates the probability that the observed results are consistent with the null hypothesis.

99	*	A value in statistics that measures significance?
100	4	A p-value is the probability, given a hypothesis, of data being as extreme as or more extreme than what is observed.
101	2	The probability or likelihood that you would find the parameter/statistic/difference that you observed given the assumption that the parameter/statistic/difference is 0.
102	*	An indicator of whether the results of a statistical test are significantly beyond an established cutoff.
103	2	It is a value that represents the probability of an event occurring.
104	*	The significance test of your results. It provides the test results of the null hypothesis and the alternative hypothesis.
105	*	Used to decide if you should reject the null hypothesis
106	1	I don't remember. (Then I googled and found that it was the probability of observing results beyond a certain threshold given the null hypothesis. I answer the following questions based on solely this understanding.)
107	2	A value by which aggregate data that falls under indicates a true difference and not occurring by chance
108	*	a number range from 0 to1 in hypothesis testing to test the significance.
109	2	The lowest value of a test level for which we would reject the null hypothesis.
110	2	the probability that a randomly selected point will be a certain value
111	2	the probability that something happened by chance
112	4	Given that a hypothesis about the value of a parameter is true, that is, under the null hypothesis, the p-value gives you the probability that the actual observed estimate, or values greater than it, would be in fact be drawn from the implied distribution under the null hypothesis
113	*	P value is a marker of significance.
114	*	If p-value is less than alpha then we reject Null Hypothesis.
115	*	used to test significance of treatment factors
116	2	p value is the significant level which you use to set up the rejection region.
117	*	A p value is a number that is calculated that can give you if your model Ian significant.
118	3	It indicates an evidence against the null hypothesis
119	2	P value is used for statistical hypothesis testing, it is probability that null hypothesis is true.
120	2	The level of significance within a statistical hypothesis test that represents the probability of an event occuring
121	*	A p-value is a measure that is used to accept or reject a null hypothesis for a formal statistical test. A threshold, alpha, is chosen which corresponds to the probability of rejecting the null hypothesis when it shouldn't have been.
122	*	a p-value determines the significance of an experiment. If the p-value is lower than the alpha value then we can reject the null hypothesis and the p-value is higher than the alpha value we fail to reject the null hypothesis

123	*	A p-value is a standard of significance
124	3	P-value is the value that indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
125	*	is a measurement that helps assess the null hypothesis, a small p-value means that we reject the null, a large p-value (as compared to the significance level) means that there is not enough evidence to reject the null hypothesis
126	3	The probability of the data falling in a certain region of the distribution when testing a hypothesis statistically.
127	3	The are under the probability curve to the right, mostly use to evaluate hypothesis
128	*	P-value is a number that helps you determine the significance of your results.
129	2	A p-value is an estimated probability to the truthfulness of the null hypothesis.
130	2	p-value is the probability for x > one particular value
131	4	probability for a given statistical model that, when the null hypothesis is true, the statistical summary would be the same as or of greater magnitude than the actual observed results
132	4	The probability of finding the observed results or potentially more extreme results, given the assumption that the null hypothesis is true
133	3	It is a probability for a given statistical model that, when the null hypothesis is true.
134	2	Datasets obtained usually fall within a range of a normal distribution. The p-value represents the probability of obtaining an observation value beyond the set threshold of Type I error (alpha = 0.05 for example). The lower the value obtained, it means you have a lower the chance of obtaining an extreme value beyond your threshold. Hence if your p-value is 0.0001, it means you have 0.01% chance of obtaining that observational value, which you can deem unlikely and thus make decisions about your test.
135	2	A number to show the correlation between few factors.
136	4	The probability of observing your results or something more extreme when the null hypothesis is true.
137	2	A p-value is a value that reflects how statistically rare a result is assuming the null hypothesis is true.
138	2	It is the likelihood of rejecting the null hypothesis of a statistical test while its actually true. The lower the value, the more significant the result is.
139	2	The probability of random fluctuations detecting an observed effect at least as large as the population in the sample.
140	2	a given statistical model that, when the null hypothesis is true
141	2	the probability of having error type 1
142	3	The probability of an observation being as extreme as the one observed in the sample collected.
143	*	a reference value to determine whether to reject or fail to reject null hypothesis

144	2	A statistical value which represents the likelihood of something being significant.
145	4	the probability of obtaining a sample which is equal to or more extreme than the data at hand under the null hypothesis
146	2	Possibility for Ho.
147	2	The likelyhood that the information explained is due to random chance.