Introduction

It is unfortunate that the best tools available for predicting job performance are often the same ones that produce the largest sub-group differences. This is a very real social problem that rests in part on the shoulders of psychologists. The search for alternative selection instruments that predict as well as cognitive ability measures without producing the same sub-group differences has been an arduous one, often fraught with disappointment (Schmidt, 1988). Strategies such as changing the context of cognitive ability measures to be more in line with cultural values (DeShon, Smith, Chan & Schmitt, 1998) and supplementing them with personality measures (Ryan, Ployhart & Friedel, 1998) have been met with limited success. Hunter & Hunter (1984) found work sample tests to have validities closest to those of cognitive ability measures. Work samples and assessment centers have proven to produce considerably less adverse impact, though often costing significantly more (roughly ten times per applicant) to develop and administer (Hoffman & Thornton, 1997).

Situational judgment tests (SJT’s) have recently shown potential to be a viable alternative selection tool. They have demonstrated predictive validity potential (McDaniel, Bruhn-Finnegan, Morgeson, Campion, & Braverman, 2001) while at the same time consistently producing lower levels of adverse impact (Pulakos & Schmitt, 1996). SJT’s are also less cumbersome relative to other simulations (i.e. works samples and assessment centers) in terms of resources necessary to develop and administer. Despite the rising interest in and use of SJT’s, the fundamental measurement properties of these instruments remain poorly understood. That is, it has yet to be determined exactly what is measured and how it is measured (Ployhart, 1999). Thus, the utility of SJT’s is limited for a number of reasons.

First, there is quite a bit of confusion about what constructs are measured or can be measured by SJT's. Logical intuition has lead some to believe that situational judgment is essentially isomorphic with Wagner & Sternberg's (1985) tacit knowledge, however limited empirical evidence suggests otherwise (Mullins & Schmitt, 1998). Similarly, others have argued that because SJT items are often multidimensional it is likely that "situational knowledge" mediates the relationship between KSA's (such as job knowledge, experience and general ability) and job performance (Motowidlo, Hanson & Crafts, 1997; Chan & Schmitt, 1997). There is scant empirical evidence to substantiate these claims.

There is also quite a bit of speculation (e.g. Chan & Schmitt, 1997; Weekley & Jones, 1999; McDaniel et al, 2001; Ployhart, 1999; Motowidlo et al, 1997), but little empirical evidence to suggest that SJT's are best conceptualized as a "hollow" measurement method rather than indicators of a unique construct or combination of constructs. If this is true, one should be able to measure any number of constructs using SJT items as the modality. Furthermore, it is unclear what proportion of the variance explained by SJT's is truly due to the constructs allegedly measured and what proportion is due to the method itself. Research in this area is necessary to gain an understanding of their predictive validity.

It is therefore crucial to step back and remind ourselves that validity evidence needs to go beyond criterion referenced studies (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education., 1999; Messick, 1995; Federal Register, 1978). Much of the rationale for designing alternatives to cognitive ability measures lies in the social issues that accompany measures of g. If I/O psychologists are to develop alternative selections methods under the banner of being socially responsible, it should be done in a proper scientific manner. From a basic science standpoint, it is not enough to demonstrate that we can create an a-theoretical measure that predicts job performance while reducing adverse impact. Selection devices should be rooted in sound theory such that those who administer them have an understanding of what they are measuring and thus a grasp of the linkage between predictor and criterion. We need to know not only that our measures predict, but also have a firm understanding how and why they do (AERA et al, 1999).

The development of SJT's is largely a-theoretical (Motowidlo, Dunnette, & Carter 1990; Motowidlo et al 1997). The repercussions of this are seen in the enormous variability in empirical findings (McDaniel et al, 2001). There is little that is common among SJT's. In fact, it seems that the only thing common among most SJT's is the way in which they are developed, thus bolstering the speculation that SJT's are nothing more than scaffolding for measuring various constructs. If one hopes come close to challenging cognitive ability measures as the best predictors of job performance, work in the realm of theory and fundamental understanding needs to be done. Otherwise we are likely to remember SJT's as a failed attempt to create a more socially responsible alternative selection tool (Schmidt, 1988).

Moreover, understanding SJT's on a conceptual level is more than just an epistemological issue. While SJT's are considerably easier to develop and administer than a full-scale assessment center, they are not an "out of the box" predictor. Their utility and generalizability is limited as a result of the fact that they remain poorly understood at the construct level. That is, SJT's appear to be useful only for the unique application for which they are designed because criterion related validity evidence is not accompanied by evidence of construct validity. Therefore there is no evidence that a SJT designed to predict performance in one application will be similarly predictive (i.e. generalize) in another application in which similar performance dimensions are relevant because there is no evidence that the SJT measures what is purports. While it is not uncommon for a test to be valid in one circumstance and less valid in another, if construct validity issues are clarified (i.e. if it can be established that SJT's measure their purported constructs), future development and administration can be facilitated. In practical and economic terms, this translates to reducing the time and energy required to develop SJT's. Ideally, one could have more generalizable individual tests, guided by theory, that are functional and predictive across many applications and would only require minor contextual modification to fit other application needs. Researchers and test developers armed with knowledge of fundamental psychological processes capable of being measured with SJT's could potentially construct tests on a more basic structural level (i.e. based on known relationships between various constructs and job performance) that would generalize to a variety of application domains. In order to reach such a goal, more research is needed in the realm of construct validity.

It has been more than ten years since Motowidlo, Dunnette and Carter's (1990) "resurrection" of SJT's or low fidelity simulations. Up until now, much of the work that has been done on the construct validity of SJT's involves little more than examining correlates of these measures. While this is useful as an exploratory technique and provides and important piece of the puzzle, more precise and informative methods are in order. The multitrait-mutimethod matrix is one such technique (Campbell & Fiske, 1959). Analyzing MTMMM data has become even more advanced and powerful with the use of confirmatory factor analysis (Kenny, 1976). CFA provides more information in that it allows the separation of method, trait and error factors (Millsap, 1990). This is ideal for examining the measurement properties of SJT's, which is essential for establishing construct validity.

The present study contributes to the literature by addressing construct validity issues of SJT's using methodologies more precise and informative than those used previously. I evaluate the extent to which SJT's represent a measurement method and also gain insight into the proportion of variance due to trait, method and error factors. The ultimate goal here is to improve the fundamental understanding of how SJT's predict. Analyzing MTMMM data appears to be the next logical step in accomplishing this goal. Thus, the present study designs a SJT of specific Five Factor Model (Goldberg, 1990; Costa & McCrae, 1992) personality traits such that the extent to which SJT's are capable of exhibiting convergent and discriminant validity can be examined. In addition, the use of CFA allows a quantifiable estimate of the extent to which the SJT method contributes to the variance explained by these measures.

Simulations

Simulations have long been established as a viable means for predicting job performance (Wernimont & Campbell, 1968). The very notion of a simulation as a personnel selection device is intuitively appealing. It allows those who administer such measures a glimpse into an applicant's potential knowledge, skills, abilities (KSA's) or dispositions and can also be focused on behaviors. The psychological literature has seen a recent renewed interest in situational judgment tests (SJT's), or what Motowidlo, Dunnette and Carter (1990) termed "low fidelity simulations." The amount or degree of "fidelity" a simulation can be described to posses depends on how directly it relates to actual job performance in terms of mundane realism. High fidelity simulations involve applicants performing a number of tasks that directly and faithfully sample from the job for which they are applying. Individuals may take part in an assessment center for a lengthy period of time and actually perform a sample of tasks taken from the job to which they are applying. Assessment centers may also be designed to measure specific KSA's relevant to a particular position. The central distinction is that an actual or closely veridical response for performing the task is made. For example, those applying for administrative positions may be asked to complete a series of filing tasks commonly encountered on the job (Palmer, Boyles, Veres & Hill, 1992).

As the fidelity of a simulation decreases, so does this element of situational realism. Low fidelity simulations involve paper and pencil or video based tests in which the individual is asked to place herself in a hypothetical situation and choose from a number of alternatives as to which she believes is the best and worst course of action. The response in the case of the low fidelity situation is a written or spoken one, as opposed to an action in a high fidelity simulation.

Similarly, neither the problem to be solved not the applicants response is real (Motowidlo et al., 1990; 1997).

As one might imagine, as fidelity decreases, so does the cost, while ease of development and administration increases. The interest in low fidelity simulations stems from this notion. The underlying motive is to keep the psychological realism intact, while reducing the need for expensive props and role-playing activities. If we can develop an SJT with most of the predictive power of a high fidelity simulation (e.g. Gaugler, Rosenthal, Thornton & Benson, 1987 report an uncorrected mean validity of .30 for assessment centers) without incurring the immense cost of development and administration that accompanies higher fidelity simulations, researchers, organizations, and practitioners can reap considerable benefits. Research tentatively suggests that predictive validity remains, for the most part, intact for low fidelity simulations or SJT’s (e.g. McDaniel et al, 2001 report a corrected mean validity of .34).

SJT’s can be roughly conceptualized as a paper and pencil or video form of an assessment center or work sample exercise in which both the problem and the response are hypothetical (Motowidlo et al, 1997). Many parallels can be drawn from past assessment center literature to the current state of the SJT literature. For example, many of the dimensions reportedly measured by assessment centers are remarkably similar to those often measured by situational judgment tests. Such dimensions include analytic skills, interpersonal skills, communication and decision-making (Sackett & Dreher 1982). In addition, much of the validity evidence for assessment centers is of the predictive and content form. It has been widely demonstrated that assessment centers predict job performance, but definitive reasons for why they predict remain elusive (Arthur, Woehr, & Maldegen, 2000). Investigations into the construct validity of assessment centers demonstrated that the variance explained by assessment centers was largely due to method or exercise factors rather than the KSA dimensions purportedly measured (Sackett & Dreher, 1982; Bycio, Alvares, & Hahn 1987; Sackett & Dreher, 1984, Brannick, Michaels & Baker, 1989). A vast majority of these researches interpreted the emergence of method or exercise factors to represent performance irrelevant method variance. Nevertheless, others (Neidig & Neidig, 1984; Lance, Newbolt, Gatewood, Foster, French & Smith 2000) argue that exercise effects represent cross-situational specificity in performance across dimensions. That is, these effects represent criterion valid task performance that should not necessarily produce construct validity in the form of trait or dimension factor loadings.

More recent attempts at construct validation of assessment centers suggest that variability in the rating process (within exercise vs. within dimension) may shed more light on findings of significant method factors. When assessors rated participants on each exercise, the typical exercise factors emerged, but when assessors rated participants across exercises on dimensions, dimension factors emerged (Robie, Osburn, Morris, Etchegaray, & Adams 2000). Similarly, Arthur et al (2000) examined the extent to which variance explained by methodological factors such as type of assessor and exercise interfered with obtaining convergent and discriminant validity using generalizability theory in addition to confirmatory factor analysis. Person (the level of agreement in ratings for an individual across exercises and dimensions) and dimension (systematic variance associated with the dimension) effects explained sixty percent of the total variance in their assessment center, while methodological effects explained only eleven percent of the total variance. These findings have lead their respective authors to conclude that assessment centers can and do exhibit convergent and discriminant (i.e. construct) validity.

Recent work on assessment centers recognizes the consistent, yet seemingly contradictory findings of almost twenty years of research. Assessment centers consistently demonstrate criterion related validity but not construct validity. Nevertheless, some empirical work (e.g. Arthur et al, 2000; Robie et at, 2000) suggests that assessment centers that are more rigorously designed are capable of overcoming this apparent contradiction. Others (e.g. Lance et al 2000) explain the contradiction by arguing that assessors are providing task rather than trait ratings. The fact that so many analogies can be drawn between assessment centers and SJT's suggests a certain level of caution before the field of I/O psychology should be willing to sanction the broad use of SJT's. It is quite possible that the lack of convergent and discriminant validity evidence is due to construct irrelevant method effects. It is also possible that the SJT method measures something other than the purported constructs, but that is nevertheless predictive of the criterion. This explanation is somewhat analogous to the cross situational specificity argument raised in the assessment center literature. Finally, it is possible that SJT's are not developed with the adequate rigor necessary to demonstrate convergent and discriminant validity.

Adverse Impact

The most compelling reason for adding SJT's to the cache of legitimate selection tools is that they have shown potential to produce significantly less adverse impact relative to otherwise extremely predictive cognitive ability measures (Hunter & Hunter, 1984; Strong & Najar, 1999; Chan & Schmitt 1997; Pulakos & Schmitt, 1996). While it may be naïve at this point to believe that these measures can eliminate adverse impact, research has consistently shown the black-white differences on SJT's to be in the neighborhood of .5 SD or less (Motowidlo & Tippins, 1993; Weekley & Jones, 1997; Weekley & Jones, 1999; Clevenger, Jockin & Morris, 1999), a substantial improvement over typical 1 SD difference associated with cognitive ability measures (Gottfredson, 1988). In addition, a number of studies have noted that SJT's are perceived by test takers to be more face valid (Motowidlo et al, 1990; Chan & Schmitt, 1997). The fact that SJT's are perceived as job-relevant has implications for test taking motivation as well as potential to avoid conflict with job applicants. Chan, Schmitt, DeShon, Clause & Delbridge (1997) demonstrated that test-taking motivation was positively related to performance on cognitive ability measures and sub-group differences were partially a function of motivation. Thus, SJT's appear to be a more socially responsible predictor than those currently available. Potential to avoid time consuming, costly and counterproductive litigation may be where gains are to be had. With this comes the need for further research and understanding of these measures.

Development of SJT's

Despite suggestions that SJT's involve considerably less commitment in terms of time and resources required of higher fidelity simulations, the current state of affairs apparent in the SJT literature suggests these measures *do* involve considerable investments. The process is often a lengthy, multistage one, which requires contributions from a large number of individuals. This translates to being expensive and cumbersome, especially in the case of video based SJT's using professional actors and editors (Dalesssio, 1994, Weekley & Jones, 1997). While this investment may be relatively small in comparison to that involved in a full-fledged assessment center, the necessary time and energy may remain prohibitively high. Employers may choose to avoid this investment and utilize established selection instruments, such as cognitive ability measures, that are easier to administer and predict more accurately (Schmidt & Hunter, 1981; Hunter, 1986). However, such measures are widely known to produce significant sub-group differences.

The current process for developing and constructing a paper and pencil or video based SJT essentially involves starting from scratch in a relatively a-theoretical manner for each unique application (Phillips, 1992, 1993; Chan & Schmitt, 1997). The traditional paradigm for SJT construction involves three stages using three independent samples. The first stage involves gathering critical incidents of performance/construct domains from an incumbent sample. This sample essentially creates the item stems or vignettes. A second independent sample reviews the critical incidents produced by the previous sample and is asked to generate general strategies for dealing with each situation. A third and final independent sample of experts identifies the most effective and least effective strategy generated by the previous sample for each item (Motowidlo et al, 1990).

It should be noted that the very nature of this development procedure is not conducive to producing tests containing unidimensional items (i.e. those designed to measure a single specific construct). While SJT developers may have a priori notions of which constructs are to be measured by the instrument (i.e. interpersonal effectiveness etc), the fact that job incumbents from diverse backgrounds who are naïve to principles of psychometrics and notions of factor-pure scales are primarily responsible for generating the item stems and responses does not give test developers a high level of control over the dimensionality of the items (Ployhart, 1999). Ployhart (1999) has suggested that this paradigm be altered such that SJT are explicitly developed to measure a priori constructs. That is, test developers should be mindful of the to be measured constructs at all stages of test development and take precautions to ensure deviations from the constructs purportedly measured are minimized.

Criterion-Referenced Validity

The majority of investigations into the validity of SJT's have been criterion related in nature. Motowidlo et al (1990) found concurrent validity estimates ranging from .28 to .37. In a 1993 extension of that research, Motowidlo and Tippins found predictive validity estimates ranging from .13 to .28. In another predictive design, Dallessio (1994) found a video based SJT to be a significant predictor of job turnover, with total score validities ranging from .12 to .29. Weekley and Jones (1997) found validity coefficients of .33 in a developmental sample (i.e. concurrent validity) and .22 in a cross validation sample (i.e. predictive validity). A second video based SJT had a predictive validity coefficient of .33. In two rather large samples, Weekley and Jones (1999) obtained validities ranging from .16 to .23 in cross validation samples. A recent meta-analysis reports a corrected population validity of .34 (McDaniel et al, 2001). Such findings suggest that SJT validities are comparable to those of assessment centers, which have an uncorrected mean validity of .30 (Gaugler et al, 1997).

A small handful of studies have examined the extent to which SJT's are predictive of job performance above and beyond already established measures. Weekley and Jones (1997, 1999) first regressed both cognitive ability and experience on job performance and then added their SJT into the regression equation. Two studies in the 1997 publication found incremental validity to be .025, .057, and .096 for three separate SJT's. The 1999 publication reported incremental validities of .033 and .011. More recently, Clevenger, Pereira, Wiechman, Schmitt and Schmidt-Harvey (2001) examined the extent to which SJT's are predictive above and beyond job experience, cognitive ability, conscientiousness and job knowledge in three samples. Clevenger et al first regressed these four variables on job performance in step one and then regressed their SJT on job performance in step two. They found the incremental validity to be significant in two of the three samples. It is evident from the limited research on incremental validity that at least some SJT's capture variance beyond that of general ability and past experience.

Nevertheless, the generalizability of criterion related validity coefficients are weakened without a firm understanding of what is measured. When we administer cognitive ability or personality measures for the purpose of prediction, we have a relatively firm understanding of what we are measuring and why it is predictive. Similarly, there are logical and theoretical linkages between predictor and criterion. Such is not the case with situational judgment tests. Many seem to take for granted, or perhaps overlook in light of criterion related validities, that SJT’s do in fact measure what is intended. If SJT’s are to be legitimate selection tools with a credible future in personnel selection, the fundamentals of what is measured and how it is measured need further exploration. At this point, all that can be said is that they do predict job performance, but how and why has yet to be determined (McDaniel et al, 2001; Pereira & Schmidt-Harvey, 1999).

Construct Validity

SJT’s are once again analogous to assessment centers in that almost all consistent validity evidence is criterion or content referenced in nature (McDaniel el al, 2001). A number of researchers (e.g. Arthur et al, 2000; Lance, et al 2000) note that validity evidence for assessment centers is paradoxical in the sense that these high fidelity simulations display promising criterion and content related validity but often fail to demonstrate acceptable levels convergent and discriminant validity. While there are fewer and less sophisticated empirical attempts to examine the construct validity of SJT’s, they appear to be similarly paradoxical. Construct validity efforts to date have involved little more than examining correlates of SJT’s (Ployhart & Erhart, 2001) and have failed to offer much in the way of meaningful or consistent insight into convergent and discriminant validity

Research attempting to examine the construct validity of SJT’s has been largely inconsistent. Constructs that are measured across all SJT’s have thus far proven elusive in that no consistent relationships have emerged with other variables such as personality and general cognitive ability (Chan & Schmitt, 1997; Mullins & Schmitt, 1998). For example, some researchers have found significant relationships with experience (e.g. *r* = .16 to .19) (Weekley & Jones, 1997, 1999; Ployhart, 1999), while others have found non-significant relationships (e.g. *r* = -.04 and .09) (Mullins & Schmitt, 1998; Bess, 2001). General cognitive ability has been among the most enduring constructs examined, but results have also been mixed—with some researchers finding significant correlations with g (e.g. *r* = .42 and .46) (Weekley & Jones, 1999; McDaniel et al., 2001) and some finding non-significant relationships (e.g. *r* = -.04 to .09) (Mullins & Schmitt, 1998; Motowidlo et al., 1990; Ployhart, 1999). McDaniel et al's 2001 meta-analysis suggests a corrected mean correlation of .46 with cognitive ability, but with enough non-trivial variability (i.e. credibility values ranging from .17 to.75) to cast doubt on any conclusions based on this data. Research in the personality domain has been sparse, but some researchers have also found relationships between SJT's and FFM personality variables such as conscientiousness (*r* =.26-.32), extraversion (*r* = .19-.20) emotional stability (*r* = .16-.19) and agreeableness (*r* = .22-.24) (Smith, 1996; Bess, 2001; Mullins & Schmitt, 1998, Smith & McDaniel, 1998).

Very few attempts have been made to *directly* examine the extent to which SJT's measure their purported constructs. Much of the work described above which has been termed construct validity or relating to constructs measured by SJT's has involved a "shotgun" like technique of correlating measures of ability or personality with SJT's without giving much thought to logical linkages between relationships. While this practice may provide broad and exploratory evidence for convergence, it has revealed the need for more precise validation techniques. If SJT's are to be widely used and accepted, test developers should have something more in the way of evidence to show that a test designed to measure, for example, interpersonal effectiveness or problem-solving effectiveness, does in fact measure these dimensions.

Furthermore, theoretical rationale needs to be provided for why hypothesized relationships with other variables exist. For example, there does not appear to be any theoretical basis to expect a relationship between a SJT that measures problem solving ability and a measure of extraversion or agreeableness.

From a measurement perspective, Chan and Schmitt (1997) have made the most diligent construct validation attempt to date. In order to establish that a video and paper and pencil SJT were measuring the same constructs, they performed a confirmatory factor analysis to test for measurement invariance. They found equal factor loadings across methods, suggesting that both forms of the SJT were indeed tapping the same thing. Unfortunately, such attempts at construct validation are the exception rather than the rule.

Ployhart (1999) has made the most thorough attempt at a construct valid SJT by explicitly building a priori constructs into the test development procedure. That is, each item was explicitly linked to a customer service dimension, which was in turn linked to a personality dimension. It was believed that this method for developing an SJT's would eliminate, or at least attenuate the multidimensional nature of SJT items and thus result in a construct valid measure. Nevertheless, the SJT used in this study did not demonstrate impressive convergent or discriminant validity ($r = -.17$ with Neuroticism, $r = .21$ with Agreeableness, $r = .22$ with Conscientiousness). Perhaps the method of analysis described later is partially responsible for this finding. In addition, the fact that personality traits were measured through customer service proxy variables that could be relevant to (and were explicitly "allowed" to) or linked to more than one FFM trait may have resulted in a multidimensional measure.

One possible reason for inconsistent findings and small convergent validity coefficients is that much of this research examines relationships between measures of fairly narrow unidimensional personality or ability constructs and SJT's designed to measure multiple constructs. That is, even SJT's specifically designed to map onto constructs such as FFM personality traits (e.g. Ployhart, 1999) do not ostensibly show strong convergent and discriminant validity because the SJT *as a whole* is correlated with or regressed onto a single trait. Rather, it seems more appropriate to examine covariations between specific SJT items or scales designed to measure a particular trait with other items or scales of that single specific trait. For example, Ployhart (1999) designed a SJT to measure personality constructs (i.e. agreeableness, conscientiousness and neuroticism) relevant to the customer service industry. Despite the fact that these dispositional constructs were built into the development of the SJT a priori, the validity analysis consisted on correlating the entire SJT with each personality construct. That is, the SJT designed to measure agreeableness, conscientiousness and neuroticism was first correlated with agreeableness, then with conscientiousness and then with neuroticism instead of correlating the specific items or scales designed to measure each trait with their appropriate counterparts. It is not surprising then that a measure of three traits shows a correlation with a single trait in the range of .2. Perhaps the items specifically designed to measure conscientiousness are highly correlated with other measures of conscientiousness and the fact that items designed to measure other traits are included in the analysis attenuates this relationship. Additionally, the majority of SJT's are developed following Motowidlo et al's (1990) procedure, which does not prescribe that individual items be explicitly linked a priori to specific traits or performance dimensions. Items are sorted into trait or performance dimension scales post hoc. This procedure is likely to result in poorly specified boundaries between performance dimension or trait scales.

Researchers have cited a number of other possible reasons (that are contradictory at times) for the immense amount of variability in SJT's relationships with other variables. Many allude to the multidimensional nature of situational judgment items (Chan & Schmitt, 1997, Vasilopoulos, Reilly, & Leaman, 2000). Chan and Schmitt (1997, p. 145) argue, "The individual situational judgment problem is nearly always multidimensional in nature in the sense that adequate solution or handling of the problem would involve several ability and skill dimensions." This implies that SJT's mediate the relationships between a combination of other KSA's such as job knowledge or cognitive ability and job performance.

Although some would like to think that judgment in complex situations represents a unique and independent construct (Sternberg, Wagner, Williams, & Hovarth, 1995), there has been quite a bit of recent speculation (Chan & Schmitt, 1997; Weekley & Jones, 1999, McDaniel et al, 2001) suggesting that SJT's are best conceptualized as a measurement *method*, with which a variety of constructs can be measured. While research findings are at this point inconsistent, one would think if situational judgment is indeed a unique construct that relationships between other common constructs would be trivial or at least consistent. Weekley and Jones (1999) have aptly suggested that SJT's may be better conceptualized as a methodology, rather than a measure tapping a unique construct. They note "studies of biodata, employment interviews, and assessment centers represent studies of the validity of the technique, not of a specific construct (Schmidt & Rothstein, 1994)." Similarly, Motowidlo et al (1997) and McDaniel et al (2001) suggest that SJT's are best conceptualized as a measurement method capable of measuring any number of constructs. That is, because of the lack of communality among SJT's and the similarity of SJT's to assessment centers (i.e. they can be conceptualized as a "low fidelity" form of an assessment center) it is perhaps best to investigate SJT's in terms of variance due to dimension (construct relevant) and method effects (construct irrelevant). Multitrait-multimethod matrices are often used for this very purpose (Leivens & Conway, 2001).

Multi-Trait Multi-Method Matrices

The multitrait-multimethod matrix (MTMMM) is an ideal procedure for examining the construct validity of SJT's. The basic idea is to measure more than one construct using more than one method. Ideally, one of the methods should be already established as a valid measure of the intended construct. Correlations among measures of the same trait using different methods (mono-trait heteromethod) are an indication of how well measures of the trait converge and are thus seen as evidence that they are measuring the intended constructs (Whitley, 1996). Once convergence has been established, it is equally important to compare correlations of different traits measured by the same method (heterotrait monomethod) with correlations of different traits measured by different methods (heterotrait heteromethod). Method variance is thought to exist when the heterotrait monomethod correlations are significantly greater than the heterotrait heteromethod correlations and to be absent when these correlations are similar (Millsap, 1990). This is particularly important in examining SJT's because of speculation that method factors exist.

Since the original conceptualization of the MTMMM (Campbell & Fiske, 1959) there have been advances in methodologies for analyzing MTMMM data (Kenny, 1975). Currently the most widely used and accepted method is confirmatory factor analysis. The CFA method is ideal for analysis of SJT's because models for each observed variable can be comprised of trait, method and random error components (Schmitt & Stultz 1986). Method variance is no longer defined as the difference between heterotrait monomenthod and heterotrait heteromethod correlations, but as the proportion of variance explained by method factors (Millsap, 1990). The importance of this distinction is evidenced by Williams, Cote & Buckley's (1989) re-analysis of Spector's (1987) MTMM data. Spector (1987) used the technique of comparing correlations described above in the analysis of 10 published MTMM studies and found little evidence of method variance. Williams et al (1989) analyzed the same data set using CFA and found method factors explained 25% of the variance in the measures analyzed.

The procedures outlined by Widaman (1985) have become somewhat of an "industry standard" for CFA. Widaman describes a systematic process of hypothesis testing that involves contrasting a number of models nested within the "best fit" model in order to examine aspects of convergent and discriminant validity. This methodology also allows one to estimate the amount of method variance within a measure.

CFA has proven to be a reliable, powerful and comprehensive technique for addressing construct validity issues in a wide variety of domains. It has been said, "Factor analysis is intimately involved with questions of validity…Factor analysis is at the heart of measurement of psychological constructs" (Thompson & Daniel, 1996, p. 198). For example, much of the research mentioned earlier on construct validity of assessment centers used CFA. Bycio et al (1987) tested a number of competing factor models for an assessment center and found that across all models, exercise or method contributed to more variance than ability and error combined. Lance, Teachout and Donnelly (1992) used Widaman's (1985) paradigm to test competing theoretical structures of criterion construct space of work samples exercises. They found that the full model that specified both ability and method factors fit the data better than more parsimonious nested models that specified either method only or ability only factor structures. While this study was able to establish convergent and discriminant validity by a series of comparisons between the full and nested models, it also found that method factors explained a significant amount of variance in the work sample exercises.

CFA has also been used for construct validation purposes in domains other than assessment centers. Vance, MaCallum, Coovert and Hedge (1988) used Widaman's (1985) procedure to analyze MTMM data in order to validate a job performance measure. They were able to demonstrate convergent and discriminant validity as well as estimate the contribution of method variance. Similarly, Harvey, Billings and Nilan (1985) challenged the construct validity of the Job Diagnostic Survey using CFA. They tested competing factor models of the JDS and found loadings on construct irrelevant method factors to be of comparable magnitude to factors relevant to JDS constructs.

While CFA is an extremely powerful method for assessing construct validity, there are limitations. The most glaring of which is the fact that CFA cannot offer immutable proof of a specified model. While few methodologies used by psychologists are impervious to such a criticism, CFA results are always subject to a particular line of attack. The possibility that another model will fit the data better is always present. One can never rule out the potential for a rival model to explain more variance in the data (Thompson & Daniel, 1996). Thus, it is extremely important for hypothesized models to be well thought out and strongly rooted in theory.

Present Study

The present research empirically examines the notion raised by Chan and Schmitt, (1997) Weekly and Jones (1999) and McDaniel et al (2001) mentioned above. The idea that SJT's are better conceptualized as a methodology is at this point speculation. Perhaps the reason for the immense variability found in the SJT literature is the fact that almost everyone is using a different SJT's measuring different constructs. If researchers X and Y do separate studies on the big five dimension of conscientiousness, their results are easily comparable because there are standardized and accepted measures of conscientiousness (for example, the NEO-PI). Researcher A's SJT may be completely different from researcher B's measure. Some researchers specify they are attempting to measure dimensions such as problem solving, communication interpersonal effectiveness, planning/organizing and motivation (Motowidlo et al, 1990; Clevenger et al, 2001) while others simply state that their instrument is a SJT (Dalessio, 1994). It is assumed that those who do not specify dimensions or sub-facets of their measure believe SJT's measure a unidimensional judgment construct. The obvious problem here is that a measure of communication and interpersonal effectiveness is going to differ dramatically from one that is supposed to measure motivation and organization or simply "judgment." Motowidlo et al (1997) note that SJT's are a measurement method used to assess a variety of constructs. It is clearly not necessary for two SJT's correlate with each other if they are measuring different constructs. It is unclear why researchers attempt to make broad, sweeping and seemingly universal statements about SJT's when individual items and instruments are rather diverse. It seems as though many are attempting to compare apples to oranges.

The only thing that seems to be constant or standard across SJT's is the method of development. The majority of SJT's found in published articles and conference papers follow the paradigm described in Motowidlo et al (1990). This would suggest that SJT's should be conceptualized as a methodology that is inherently construct free—not as a measurement of a unique construct (Sternberg et al, 1995) or a culmination of other variables such as cognitive ability and experience (Clevenger et al, 2001). SJT's may be nothing more than an elaborate item production rule. That is, any test containing items that follow the format of containing a short vignette describing a hypothetical situation followed by four or five potential courses of action can theoretically be called a SJT (Motowidlo et al, 1997). Thus, the test developer or researcher is free to measure any number of constructs using this methodology. This line of reasoning obviates the underlying cause of “mixed” or inconsistent findings in construct validation attempts.

Mixed findings aside, little effort has been made in the way of confirming that SJT’s do in fact measure what they say they do. It is unclear if one can effectively measure the intended performance dimensions or constructs. It is entirely possible that the variance observed in SJT’s is due to method effects rather than true variance in the traits or dimensions purportedly measured. It is also possible that criterion referenced validity remains intact because what appears to be a method effect may actually be criterion related variance attributable to unintended constructs being measured. Either way, it is essential for the viability of SJT’s that construct validation efforts be made. Without construct validity evidence, generalizability of these measures is severely limited in the sense that it will be difficult to link SJT’s to criteria related to job performance (Ployhart, 1999).

The present study examines an SJT from a MTMMM perspective by developing a SJT to measure three FFM personality traits under highly controlled conditions. More specifically, a SJT will be created in an effort to form construct “pure” scales of conscientiousness, agreeableness and openness to experience. Therefore a scale specifically designed to measure conscientiousness and only conscientiousness should converge with existing valid measures of this trait (i.e. the IPIP) and be appropriately distinguishable from (i.e. exhibit discriminant validity) measures of a separate trait such as agreeableness. Furthermore, if SJT’s are nothing more than a measurement method capable of measuring any number of constructs, SJT method factor should not explain significantly more variance than the non-SJT method factor. Similarly, trait factors should explain significantly more variance than method factors if the SJT method is truly inert. This leads to the following hypotheses:

Hypothesis 1: SJT scales created to measure specific FFM personality traits will load on the same factor as existing measures of the same personality traits (i.e. exhibit convergent validity).

Hypothesis 2: SJT scales created to measure specific FFM personality traits will load on different factors as existing measures of the different personality traits (i.e. exhibit discriminant validity).

Hypothesis 3: The SJT method factor will not have significantly larger loadings than the non-SJT method factor.

Hypothesis 4: Trait factors will have significantly larger loadings than method factors.

Method

Participants

Participants were 283 undergraduate students recruited from the introduction to psychology participant pool and other psychology courses. All participants received extra credit points towards fulfilling their class requirement. Thirty participants were removed from the sample as a result of providing either incomplete or suspect (i.e. very low variance or completing the IPIP in an unreasonably short amount of time) responses, leaving 253 participants with usable data. Of these 253 participants, 22 were African American, 27 were Asian, 190 were Caucasian, 4 were Hispanic and 10 indicated "other" ethnicity. One hundred seventy eight of the participants were female.

Measures

Personality dimensions of conscientiousness, agreeableness and openness were measured using selected items from the full 300 item version of the International Personality Item Pool

(Goldberg, 1999; IPIP 2001) designed to measure the same 5 broad facets and 30 sub-facets as the NEO-PI-R (Costa & McCrae, 1985). Relevant scales are presented in appendix A—items marked with an asterisk were used in the present analyses. Items present a short statement (e.g. "get chores done right away") and participants are asked to rate how accurately the statement describes him or her on a 5-point likert scale (1=very inaccurate to 5=very accurate). The five-factor model shows convergent and discriminant validity across different personality measures and observers (Costa & McCrae, 1992, Goldberg, 1990).

The SJT was created specifically for this study following a variation of the traditional paradigm found in Motowidlo et al (1990). The SJT was designed to measure the same three dimensions as the personality instrument above. The traditional three-step procedure was curtailed by rationally writing situations that involve and responses that reflect dimensions of conscientiousness, agreeableness and openness. That is, each item was directly linked to a specific personality trait by writing responses to reflect varying degrees of the trait described in each of the corresponding items from the IPIP. For example, the conscientiousness item found in *appendix A* (item 9) was based on the IPIP item "start tasks right away." Item responses were written on a continuous (1-4) scale such that a response of 1 indicated a low level of the trait of interest and a response of 4 indicated a high level of the trait of interest (items were written in the opposite direction as well and reverse coded to eliminate potential response bias). A total of 45 (15 for each scale) items were generated by the first author with the help of research assistants.

As an additional check to ensure that all items map onto the proper construct scale, 14 graduate student subject matter experts familiar with FFM personality dimensions were asked to sort each item into one of three trait scales or indicate that the item did not belong in any scale. 18 of the original 45 items were retained for the final SJT based on a criterion of at least 86%

agreement (i.e. 12 out of 14 SME's agreed that the item measured the appropriate trait). The entire SJT is found in Appendix B—items marked with an asterisk were used in the present analyses. Data from the non-asterisk items were excluded from the analyses because of the paths added to the model by using all or most of the items would result in prohibitively high sample size required to test such a model. Items were selected based on item intercorrelations (i.e. the 4 items with the highest intercorrelation were selected for each scale).

Procedure

All participants completed the paper and pencil SJT in a classroom setting. Participants had unlimited time to complete the measure, although most took approximately 30 minutes to finish. All participants completed the IPIP via an online test administration server. Online administration was also not under any time restriction, although most took approximately 35 minutes to finish. The order in which participants receive each measure was counterbalanced such that approximately half the participants completed the SJT first and half completed the IPIP first. A minimum of 24 hours passed in between testing sessions.

Factor models tested

The first model the null model (see figure 1), in which no common factors exist and observed variables are explained only by random error (i.e. unique variance). The null model is the most restricted model, from which restrictions can be progressively relaxed if models fail to demonstrate acceptable levels of fit (Widaman, 1985). The trait model (model 2) postulates that each observed variable loads on its intended trait, thus all variance is explained by oblique traits and random error (see figure 2). The third model (method model) postulates that only oblique method factors exist, and therefore all SJT items load on one method factor and all IPIP items load on a second method factor (see figure 3). The fourth model is a general trait model in which all observed variables are explained by a general trait factor, their respective oblique method factors and random error (see figure 4). The fifth model (see figure 5) contains oblique trait factors, orthogonal method factors and unique variance. The final model in the formal analyses, (model six) is the full trait and method model, in which variance in the measures is explained by respective oblique trait and method factors as well as error variance. That is, both SJT and IPIP items load on their appropriate trait factors while SJT's items have loadings on a SJT method factor and IPIP items have loadings on an IPIP method factor (see figure 6).

While not part of the formal analyses or represented in Widaman's (1985) procedure, models seven and eight are included for "exploratory" or supplemental purposes. Model seven is a variation of the full model, in which SJT items are explained by an oblique method factor and random error while IPIP items are explained by both oblique trait factors, oblique method factors and random error (see figure 7). This model is included to gain a more precise understanding of the contribution of the SJT method. While it is reasonable to assume that the IPIP method does not account for a great deal of variance (i.e. the IPIP have previously demonstrated acceptable levels of construct validity), the same assumption is not warranted for SJT's. Model eight is also a variation of the full model in which SJT items are explained by a general trait factor, an oblique method factor and error variance while NEO items are explained by their appropriate respective oblique trait factors, an oblique method factor and error variance (see figure 8).

Analyses

All confirmatory factor models (see figures 2 – 8) were tested using the CALIS procedure in SAS (SAS Institute, 2001) using maximum likelihood estimation. Analyses were done on covariance matrices. Cudeck, (1989) notes that covariance structure analyses done on correlation matrices can potentially result in slightly incorrect results. Convergence criterion was satisfied for models 2 through 8. Table 1 contains the correlation matrix of manifest variables and table 2 contains descriptive statistics for each item as well as its standardized factor loading on both trait and method factors for model 6. Table 3 contains selected fit statistics for models 2 through 8, as well as χ^2 values for all models. All χ^2values in table 3 are significant at the .001 level, however χ^2statistics are sensitive to sample size and will frequently be significant even when the model fits well (James, Mulaik & Brett, 1982). Values above .90 for the non-normed fit index (NNFI; Bentler & Bonett, 1980), comparative fit index (CFI; Bentler, 1989) and goodness of fit index (GFI; Joreskog & Sorbom, 1985) indicate acceptable fit. Values of the root mean square residual (RMR; Joreskog & Sorbom, 1985) below .07 indicate acceptable fit and values below .05 indicate good fit. Values of the root mean squared error of approximation (RMSEA; Browne & Cudeck, 1993) below .06 indicate acceptable fit. As seen in table 3, fit statistics indicate that model 6 does demonstrate adequate levels of fit.

Results

Assuming model 6 provides adequate fit, convergent validity can be assessed via comparison of models 3 and 6, thereby testing hypothesis 1. This involves comparing the fit indices of a model with trait and method factors (model 6) with a model with no trait factors but the same method factors (model 3). If the measures exhibit convergent validity, trait factors will exist and model 6 will fit the data significantly better than model 3. Table 4 indicates that the χ^2difference between model 6 and model 3 is statistically significant and table 3 indicates a substantial increase in fit statistic values, thus supporting hypothesis one. Discriminant validity can be assessed via comparison of models 4 and 6, thereby testing hypothesis 2. That is, if there is discrimination among traits, a model with three traits and two methods (model 6) should fit the data significantly better than a model with one general trait and the same method factors (model

4). Table 4 indicates that the χ^2 difference between model 6 and model 4 is statistically significant and table 3 indicates a substantial increase in fit statistic values, thus supporting hypothesis 2.

Method variance can be assessed via comparison of models 2 and 6. If a model with trait and method factors fits the data significantly better than a model with the same trait factors but no method factors, it is reasonable to conclude that significant method factors exist. The extent to which the methods covary can be assessed via comparison between models 5 and 6. Table 4 indicates that significant method factors exist and there is a significant amount of covariation between them. In addition, table 3 indicates a notable difference between fit statistic values of models 2 and 6 and a small difference between models 5 and 6. This is consistent with the estimated correlation between the method factors found in table 5.

Factor loadings are found in table 2. It appears that the average and median method loading for SJT variables are roughly equivalent to the average and median method loading for the IPIP variables, thus providing support for hypothesis 3 (it should be noted that trait and method factors are likely on different scales and thus are not directly comparable—these results should therefore be interpreted as general trends rather than precise estimates). Finally, the average trait loading for SJT and IPIP items exceeds that of the method loading. Nevertheless, there is a much larger difference between IPIP trait and method loadings than for SJT trait and method loadings--indicating that the SJT method factor has non-trivial loadings on the construct irrelevant method factor in relation to construct relevant trait loadings. Additionally, in 8 out of 12 cases the SJT method loading is either larger than or has an overlapping (95%) confidence interval with the trait loading whereas the same is true for the IPIP in only 4 out of 12 instances (see table 2). These results taken together fail to, or provide only tenuous support for hypothesis 4 with respect to the SJT.

Fit statistics for the supplemental models (i.e. models 7 and 8) are found in table 3. Fit indices indicate that model 6 fits substantially better than model seven and somewhat better than model 8. These analyses provide additional support for the convergent (6 vs. 7) and discriminant (6 vs. 8) validity of the SJT.

Table 6 contains correlations among latent variables. The hypothesized model from which these relationships are derived is one with six latent trait factors, one for each construct measured by each method. IPIP trait factors were estimated using the full 10 item scales, whereas, the SJT trait factors were estimated using the 4 item scales. That is, a trait factor for conscientiousness defined by 10 IPIP items and a trait factor for conscientiousness defined by the 4 SJT items was hypothesized. Similar trait factors were hypothesized for openness to experience and agreeableness within the same model. The purpose of this final analysis was to examine the extent to which latent traits measured by each method are interchangeable or how well traits converge across methods. As indicated by table 6, correlations that represent convergence among traits (e.g. openness measured with the IPIP with openness measured with the SJT) are larger than all others, which is to be expected because each latent trait is presumably correlating with itself. Nevertheless, the absolute magnitude of each (with the possible exception of openness) suggests latent traits measured with each method are not isomorphic.

Discussion

Results of this study indicate that SJT's are capable of demonstrating construct validity in the form of convergent and discriminant validity. In addition, results suggest that significant method factors exist, at least in the present analyses. Trait loadings are considerably larger than method loadings for the IPIP variables, but the same is not true for SJT items. Although the separate method factors are not directly comparable, this can be viewed as a rough index that the method factors are explaining a larger proportion of the total variance in SJT items than in IPIP items.

It can be concluded based on the results of this study that SJT's can indeed function as a modality for measuring any number of constructs, with the caveat that the method itself may explain substantial proportions of construct-irrelevant variance relative to the proportion of variance explained by KSAO constructs one wishes to measure. Overall the study and results provide support for the conceptualization of SJT's as a measurement modality rather than indicators of a new and unique construct, although it is not outside the realm of possibility that the SJT method factor is actually some kind of "judgment" construct. The fact that the SJT method factor has a substantial relationship with the IPIP method factor (see table 5) argues against this explanation and is more consistent with conceptualizing the method factors as systematic construct irrelevant variance.

Arthur et al (2000) note that assessment centers are primarily a method of measurement and thus draw a distinction between the means of measuring constructs and the content being measured. It seems reasonable that a similar distinction is warranted with SJT's. This is likely to be the cause of the enormous variability found in SJT's studied in McDaniel et al's (2001) meta-analysis. That is, the SJT's examined in the meta-analysis measure a wide variety of content using the same modality of measurement. Just because they are fall under the rather nebulous category of "SJT" does not speak to the content each attempts to measure. The fact that the modality is the same does not mean that variety of latent constructs measured should show reliable correlations with one another or with other constructs such as cognitive ability or personality. Just as one would have little reason to expect latent constructs measured in an assessment center designed to evaluate CEO's to correlate with ones designed to evaluate clerical workers, one has equally little reason to expect divergent constructs measured with SJT's to correlate with one another or external criterion.

It is likely that previous attempts to find construct valid SJT's have failed because the traditional process of constructing SJT's often produce multidimensional items. While such tests are certainly useful for omnibus prediction purposes, no specific statement of relations between predictor and criterion (e.g. Jane is high on "interpersonal effectiveness" and thus would make a good customer service representative) are warranted. Those who use SJT's for selection purposes should have evidence that SJT scales measure what is claimed in order to have a theoretical grasp of the relationship between predictor and criterion. It is important to know what one in measuring because the ideal purpose of SJT's as an "alternative selection" tool is to measure predictive constructs other than general cognitive ability. McDaniel et al's (2001) meta-analysis indicates that there is wide variability in the *g* loading of SJT's. Convergent and discriminant validity evidence of dimensions measured by SJT's would be of great benefit in order to be sure one is not inadvertently measuring *g* and thereby defeating the overarching purpose of the alternative selection measure.

An alternate explanation for thus far failed attempts at finding construct validity evidence for SJT's is rooted in the "criterion problem" (Austin & Villanova, 1992). That is, many SJT's may in fact be construct valid but no attempt to validate the specific measured constructs was made. Much of the research examining construct validity of SJT's involves little more than correlating SJT's designed to measure problem solving or interpersonal effectiveness with established measures of cognitive ability or personality (Ployhart, 2000). Convergent and discriminant validity evidence may emerge in existing SJT's if more direct criterion measures of the specific dimensions are used.

Level of rigor in development is likely to have an effect on the convergent and discriminant validity of SJT. The SJT in the present study was afforded exceptionally high levels of developmental rigor because of the existing robust convergent and discriminant validity evidence of the FFM. Developmental rigor as an explanation for convergent and discriminant validity evidence (or lack thereof) seems to have found support in the assessment center literature. In a meta-analysis of 34 MTMMM assessment center studies, Lievens and Conway (2001) found that assessment centers that were more rigorously developed were more likely to exhibit adequate dimension factors. In addition, they found that psychologist assessors provided a better measurement of construct relevant dimensions than did managers. They conclude that psychologists' extensive training, education and knowledge of individual difference variables are likely the cause of this finding. SJT's often use job incumbents and non-psychologist SME's to develop item vignettes and stems. It is reasonable to believe that reliance on individuals who do not have extensive knowledge of individual difference variables or psychometric principles to develop dimension scales will result in a weaker measure with regard to convergent and discriminant validity.

Limitations

The SJT used in this study is admittedly artificial and contrived. That is, the majority of existing SJT's are not designed with a heavy emphasis on construct validity in mind because many use them for applied purposes and are thus more concerned with criterion related validity. Nevertheless, almost all studies involving SJT's reviewed by the author claim their measure is designed to tap specific KSAO dimensions. It is reasoned that the SJT described in this study should be considered as something analogous to a laboratory instrument, and is designed to uncover the underlying mechanisms (i.e. method, trait and error) that compose the SJT paradigm. External validity of this study may be an issue for some. Once can certainly argue that the SJT described in this study is not representative of those used in practice. More traditional SJT's typically have right and wrong answers (or "best" and "worst"), while the present SJT measured behavioral tendencies that are less objective. While both traditional SJT's and the one used in this study ask "what would you do" in a give situation, the criterion of interest for traditional SJT's is "do you know the best way to handle this situation," whereas the criterion of interest in the present SJT less evaluative. Nevertheless, the functional properties and method of development of this SJT appear analogous to those developed under more traditional conditions to the extent that generalizations from this SJT to less contrived measures are appropriate (Mook, 1983). Although the items are developed under controlled circumstances, one can argue that the end product could very well have resulted from the traditional paradigm described in Motowidlo et al (1990).

Another limitation of the present study is that the SJT items designed to measure openness did not perform well. In hindsight, this is perhaps not surprising because openness is the most poorly understood and least agreed upon construct in the FFM paradigm (Hogan, 1991). Openness to experience was originally chosen over extraversion because the items were more action oriented and thus more conducive to writing SJT items. If one looks only at the SJT conscientiousness items, it appears that the SJT performs equally as well as the IPIP. Thus, future research in the realm of replication with more refined scales is warranted.

One potential reason for the SJT's lower trait loadings is that SJT items were tied to a particular context, while IPIP are context free and thus more general. For example, one who endorses the IPIP items “like to try new foods” or “complete tasks right away” likely responds in a heuristic sense according to how he or she behaves in a number of different situations. The SJT items were designed to tap the same dimensions as above, but because of the nature and format of SJT’s must do so in a specific context or situation (see appendix B). One who generally likes to try new foods or completes tasks right away may for one reason or another not choose to behave that way in a specific context. Because only four SJT items were used to measure each trait, it is unlikely that the behavioral domain was adequately sampled to capture an individuals general standing on any one particular trait. In addition, the reader should be cautioned that because the 4 items for each scale were selected based on intercorrelation (i.e. the 4 items that produced the highest α); the present scales likely capitalize on sample specific correlations among items. Sample size limitations prohibited using a cross validation procedure.

The present sample is comprised completely of undergraduate students. This is not necessarily a weakness because the purpose here is to validate the internal structure or measurement properties of the SJT used and no relationships to external criterion were hypothesized. More importantly, Smith, Hanges & Dickson (2001) found evidence for invariance of factor pattern, invariance of factor loading, and invariance of factor variance and covariance in a FFM measure across student job incumbent and job applicant samples using multi-group confirmatory factor analysis. Thus, there is little reason to believe that findings in the student sample should not generalize to job applicants or job incumbents.

Conclusion

Despite concerted efforts, (e.g. Ployart, 2000) SJT’s have until now failed to show clear evidence of convergent and discriminant validity. The present study demonstrates that it is indeed *possible* for SJT’s to measure specific constructs and provides an example of a procedure to increase the likelihood of construct "pure" SJT scales. With the proper developmental rigor, it seems reasonable that similar procedures for constructing narrowly focused SJT scales can generalize to more realistic SJT's designed to measure traditional performance dimensions. It is unclear if construct valid scales will come at the expense of predictive validity of SJT's, but there is no reason to suspect one cannot develop a SJT that is valid in both a construct and criterion-referenced sense. An ideal and truly construct valid SJT is one that demonstrates convergent and discriminant validity while at the same time exhibiting theoretically meaningful relationships with external criterion. Despite evidence for convergent and discriminant validity, one cannot conclude that the particular SJT created for this study is truly construct valid in a unified sense because of the existence of non-trivial amounts of method variance and lack of empirical evidence of criterion referenced validity. Future research should therefore expand on the present research with the goal of satisfying all criterion of the unified view of construct validity (e.g. Messick, 1995).

The SJT in this study was created under ideal or at least more favorable and controlled conditions than one would expect in a more applied settings. Thus, the fact that method factors explained such a high proportion of variance in the SJT should be of particular concern. Construct irrelevant or method variance is troublesome because it interferes with a test takers ability to demonstrate his or her competence on the dimension of interest. Validity is ultimately an issue of the inference made from a test score. When SJT's are used for prediction purposes, method variance could very well result in invalid scores or judgments about one's competency because the measure contains *reliable* variance, which is irrelevant to the construct of interest (Messick, 1995). Those who wish to make decisions based on the specific performance domain measured by a SJT should therefore be wary without direct evidence of convergent and discriminant validity as well as an estimation of the construct relevant variance measured.

The present study has ultimate value in that it demonstrates what *can* occur when measurement aspects are isolated, albeit in way distinct from what commonly occurs in practice (Mook, 1983). A solid theoretical foundation, which in this case appears to be conceptualizing SJT's as a modality of measurement, is a useful place from which to begin more advanced and informative research in this area. In light of this, it is recommended that future studies of construct validity of SJT's in addition to searching for diverse forms of construct validity focus on individual test level validity rather than searching for constructs common to all SJT's.

References

Adoption by Four Agencies of Uniform Guidelines on Employee Selection Procedures, 166, Volume 43. *Federal Register* (1978).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, D.C: American Psychological Association.

Arthur, W., Woehr, D. J. & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment center construct related validity paradox. *Journal of Management, 26*, 813-835.

Austin, J.T. & Villanova, P. (1992). The criterion problem. *Journal of Applied Psychology, 77*, 836-874.

Bess, T.L. (2001). Exploring the dimensionality of situational judgment: Task and contextual knowledge. Unpublished master's thesis, Virginia Polytechnic Institute and State University.

Brannick, M.T., Michaels, C.E., & Baker, D.P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology, 74*, 957-963.

Browne, M.W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp.136-162). Newbury Park, CA: Sage.

Buckley, M.R., Cote, J.A., & Comstock, S. M. (1990). Measurement errors in the behavioral sciences: The case of personality/attitude research. *Educational and Psychological measurement, 50*, 447-474.

Bycio, P., Alvares, K.M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology, 72*, 463-474.

Chan, D. & Schmitt, N. (1997). Video based versus paper-and-pencil methods of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143-159.

Clevenger, J, Jockin, T. & Morris, S. (1999). A situational judgment test for engineers: Construct and criterion related validity of a less adverse alternative. Paper presented at the 14th Annual Conference of the Society for Industrial Organizational Psychology, Atlanta, GA.

Clevenger, J., Pereira, G.M., Wiechmann, D. W., Schmitt, N., & Harvey, V.S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410-417.

Costa, P. T. & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13*, 652-665.

Costa, P. T. & McCrae, R. R. (1995). Solid ground in the wetlands of personality: A reply to Block. *Psychological Bulletin, 117*, 216-220.

Costa, P.T. & McCrae, R.R. (1995). Primary traits of Eyesenck's PEN system: Three and five factor solutions. *Journal of Personality and Social Psychology, 69*, 308-317.

Costa, P. T. & McCrae, R. R. (1985). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual.* Odessa, FL: Psychological Assessment Resources, Inc.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105*, 317-327.

Dalessio, A.T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology, 9*, 23-32.

DeShon, R.P., Smith, M.R., Chan, D., & Schmitt, N. (1998). Can racial differences in cognitive test performance be reduced by presenting problems in a social context? *Journal of Applied Psychology, 83*, 438-451.

Gaugler, B. B., Rosenthal, D. B., Thornton, G.C., & Bentson, C. Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511.

Goldberg, L.R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26-34.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe*, Vol. 7 (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R. (in press). The comparative validity of adult personality inventories: Applications of a consumer-testing framework. In S. R. Briggs, J. M. Cheek, & E. M. Donahue (Eds.), *Handbook of Adult Personality Inventories.*

Gottfredson, L.S. (1988). Reconsidering fairness: A matter of social and ethical priorities. *Journal of Vocational Behavior, 33*, 293-319.

Harvey, R.J., Billings, R.S., & Nilan, K.J. (1985). Confirmatory factor analysis of the Job Diagnostic Survey: Good news and bad news. *Journal of Applied Psychology, 70*, 461-468.

Hoffman, C.C., & Thornton, G.C. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology, 50*, 455-470.

Hogan, R. (1991). Personality and personality measurement. In M.D. Dunnette & L.M. Hough (Eds) *Handbook of Industrial Organizational Psychology, 2,* 874-909. Palo Alto, CA, Consulting Psychologists Press.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

International Personality Item Pool (2001). A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences (http://ipip.ori.org/). Internet Web Site.

Joreskog, K.G. & Sorbom, D. (1985). *LISREL VI, Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares*. Uppsala: University of Uppsala.

Kenney, D.A. (1975). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology, 12*, 247-252.

Lance, C.E., Newbolt, W. H., Gatewood, R.D., Foster, M.R., French, N.R., & Smith, D.E. (2000). Assessment center exercise factors represent cross situational specificity, not method bias. *Human Performance, 13*, 323-353.

Lance, C.E., Teachout, M.S., & Donnelly, T.M. (1992). Specification of the criterion construct space: an application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology, 77*, 437-452.

Lievens, F., & Conway, J.M. (2001). Dimension and exercise variance in assessment center scores: A large scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86*, 1202-1222.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A, & Braverman, E.P. (2000). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730-740.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Millsap, R.E. (1990). A cautionary note on detection of method variance in multitrait-multimethod data. *Journal of Applied Psychology, 75*, 350-353.

Mook, D.G. (1983). In defense of external validity. *American Psychologist, 38*, 379-387.

Motowidlo, S. J., Dunnette, M.D., & Carter, G.W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640-647.

Motowidlo, S.J., Hanson, M.A., & Crafts, J.L. (1997). Low fidelity simulations. In D.L. Whetzel & G.R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp.241-260). Palo Alto, CA: Davies-Black Publishing.

Motowidlo, S.J., & Tippins, N. (1993). Further Studies of the low-fidelity simulation in the form of situational inventory. *Journal of Occupational and Organizational Psychology, 66*, 337-344.

Mullins, M.E., & Schmitt, N. (1998). Situational judgment testing: Will the real constructs please present themselves? Paper presented at the 13th Annual Conference of the Society of Industrial Organizational Psychology, Dallas, TX.

Neidig, R.D., & Neidig, P.J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69*, 182-186.

Palmer, C.I., Boyles, W.R., Veres, J.G., & Hill, J.B. (1992). Validation of a clerical test using work samples. *Journal of Business and Psychology, 7*, 239-257.

Pereira, G.M. & Harvey V.S. (1999) Situational judgment tests: Do they measure personality, performance, or both? Paper presented at the 14th Annual Conference of the Society for Industrial Organizational Psychology, Atlanta, GA.

Phillips, J.F. (1992). Predicting sales skills. *Journal of Business and Psychology, 7*, 151-160.

Phillips, J.F. (1993). Predicting negotiation skills. *Journal of Business and Psychology, 7*, 403-411.

Ployhart, R.E. & Ehrhart, M.G. (2001). Effects of response instructions of the criterion related validity, construct validity, and reliability of situational judgment tests. Paper presented at the Annual Conference of the Society for Industrial Organizational Psychology, San Diego, CA.

Ployhart, R.E. (1999). An interactionist approach to assessing personality in work contexts: Construct validation of a predictor of customer service performance. Unpublished doctoral dissertation, Michigan State University.

Pulakos, E. D. & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects of criterion-related validity. *Human Performance, 9,* 241-258.

Robie, C., Osburn, H.G., Morris, M.A., Etchegaray, J.M., & Adams, K.A. (2000). Effects of the rating process on the construct validity of assessment center dimension evaluations. *Human Performance, 13*, 355-370.

Ryan, A.M., Ployhart, R.E., & Friedel, L.A. (1998). Using personality testing to reduce adverse impact: A cautionary note. *Journal of Applied Psychology, 83*, 298-307.

The SAS System for Windows. (2000). Version 8.2. SAS Institute Inc., Cary NC.

Sackett, P.R. & Dreher, G.F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67*, 401-410.

Sackett, P.R. & Dreher, G.F. (1984). Situation specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. *Journal of Applied Psychology, 69*, 187-190.

Schmidt, F.L. (1988). The problem of group differences on ability test scores in employment selection. *Journal of Vocational Behavior, 33*, 272-292.

Schmitt, N., & Stults, D.M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*, 1-22.

Smith, D. B., Hanges, P.J., & Dickson, M.W. (2001). Personnel selection and the five factor model: Reexamining the effects of applicants frame of reference. *Journal of Applied Psychology, 86*, 304-315.

Smith, K.C, McDaniel, M.A. (1998). Criterion and construct validity evidence for a situational judgment measure. Paper presented at the 13th Annual Conference of the Society for Industrial Organizational Psychology, Dallas, Texas.

Sternberg, R.J., Wagner, R.K., Williams, W.M. & Horvath, J.A. (1995). Testing common sense. *American Psychologist*, 912-926.

Strong, M. & Najar, M. (1999). Situational judgment versus cognitive ability tests: Adverse impact and validity. Paper presented at the 14th Annual Conference of the Society for Industrial Organizational Psychology, Atlanta, GA.

Thompson, B. & Daniel, L.G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 197-208.

Vance, R.J., MacCallum, R.C., Coovert, M.D., & Hedge, J.W. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. *Journal of Applied Psychology, 73*, 74-80.

Vasilopoulos, N.L., Reilly, R.R., & Leaman, J.A. (2000). The influence of job familiarity and impression management on self-report measure scale scores and response latencies. *Journal of Applied Psychology, 85*, 50-64.

Weekley, J.A., & Jones, C. (1997). Video based situational testing. *Personnel Psychology, 50*, 25-49.

Weekley, J.A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52,* 679-700.

Wernimont, P. & Campbell, J. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372-376.

Widaman, K.F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1-26.

Williams, L.J., Cote, J.A., & Buckley, M.R. (1989). Lack of method variance in self reported affect and perceptions at work: Reality or artifact? *Journal of Applied Psychology, 74*, 462-468.

Table 1

Correlations Among Manifest Variables

	Label	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	SJTC1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2	SJTC2	0.39		--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
3	SJTC3	0.36	0.39	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
4	SJTC4	0.28	0.38	0.43	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
5	SJTA1	0.05	0.03	0.00	-0.05	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
6	SJTA2	-0.05	-0.01	-0.07	-0.01	0.22	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
7	SJTA3	0.00	0.06	0.07	0.02	0.29	0.22	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
8	SJTA4	-0.11	-0.04	-0.01	-0.06	0.25	0.40	0.25	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
9	SJTO1	0.00	0.03	0.07	0.11	-0.12	-0.11	0.05	-0.17	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
10	SJTO2	-0.05	0.02	-0.13	0.04	-0.05	-0.06	0.02	-0.15	0.13	--	--	--	--	--	--	--	--	--	--	--	--	--	--
11	SJTO3	0.06	0.08	-0.04	-0.02	-0.14	-0.18	-0.02	-0.19	0.19	0.25	--	--	--	--	--	--	--	--	--	--	--	--	--
12	SJTO4	0.05	-0.09	-0.05	-0.01	-0.08	-0.03	-0.09	-0.02	0.03	0.19	0.16	--	--	--	--	--	--	--	--	--	--	--	--
13	IPIPC1	0.29	0.30	0.41	0.38	-0.01	-0.14	-0.01	-0.03	0.14	0.10	0.00	0.11		--	--	--	--	--	--	--	--	--	--
14	IPIPC2	0.31	0.30	0.44	0.34	-0.06	-0.19	-0.02	-0.07	0.17	0.05	-0.04	0.09	0.73		--	--	--	--	--	--	--	--	--
15	IPIPC3	0.31	0.31	0.40	0.28	0.09	-0.06	-0.02	0.02	0.11	-0.03	-0.02	0.02	0.53	0.51		--	--	--	--	--	--	--	--
16	IPIPC4	0.33	0.33	0.46	0.31	0.13	-0.05	0.05	-0.05	0.06	-0.01	-0.01	0.08	0.54	0.54	0.63	--	--	--	--	--	--	--	--
17	IPIPA1	0.12	0.17	-0.02	0.08	0.19	0.06	0.26	0.17	0.10	0.21	0.11	0.01	0.13	0.16	0.07	0.06	--	--	--	--	--	--	--
18	IPIPA2	0.04	0.14	0.09	0.19	0.16	0.15	0.19	0.16	-0.12	0.07	-0.04	-0.04	0.03	0.04	0.08	0.02	0.32	--	--	--	--	--	--
19	IPIPA3	0.15	0.11	0.03	0.10	0.13	0.19	0.31	0.27	-0.05	-0.01	0.00	0.11	0.08	0.11	0.06	0.03	0.34	0.33	--	--	--	--	--
20	IPIPA4	0.14	0.11	0.09	0.14	0.21	0.17	0.26	0.15	-0.07	0.02	-0.01	0.08	0.06	0.10	0.06	0.06	0.26	0.45	0.37	--	--	--	--
21	IPIPO1	0.02	-0.06	0.00	-0.01	-0.15	-0.13	0.01	-0.11	0.14	0.19	0.17	0.26	0.18	0.16	-0.01	0.04	0.10	-0.07	0.06	0.12	--	--	--
22	IPIPO2	0.07	0.04	0.09	0.08	-0.23	-0.07	-0.03	-0.11	0.15	0.19	0.24	0.23	0.17	0.18	0.01	0.03	0.04	0.02	0.02	0.09	0.43	--	--
23	IPIPO3	0.09	0.07	0.09	0.05	-0.26	-0.03	-0.01	-0.05	0.11	0.13	0.27	0.24	0.20	0.15	0.05	0.04	-0.01	-0.03	0.01	0.06	0.42	0.74	--
24	IPIPO4	0.00	0.08	0.02	0.12	-0.20	-0.02	-0.07	-0.10	0.25	0.21	0.24	0.14	0.14	0.12	0.05	0.13	-0.06	-0.13	-0.03	-0.02	0.25	0.35	0.37

Table 2

Descriptive Statistics for Items and Standardized Factor Loading Estimates for Model 6

Item	Mean	SD	Trait loading	Std Error	Method loading	Std Error	Combined r^2	Trait/Method loading 95% CI Overlap
SJT Variables								
Conscientiousness								
α = .70								
1	1.98	.80	.49***	.0673	.22**	.0800	.29	yes
2	2.10	.98	.54***	.0665	.20*	.0795	.33	no
3	1.43	.81	.69***	.0641	.12	.0776	.49	no
4	2.04	.86	.52***	.0666	.23**	.0795	.33	no
Agreeableness								
α = .60								
1	1.43	.81	.34***	.0717	.29***	.0802	.20	yes
2	1.64	1.22	.31***	.0715	.50***	.0808	.34	--
3	2.40	.95	.47***	.0796	.16*	.0702	.24	no
4	.73	.88	.37***	.0705	.54***	.0810	.42	--
Openness								
α = .43								
1	1.67	1.09	.11	.0676	.28***	.0816	.09	--
2	2.58	.74	.15*	.0671	.30***	.0812	.11	--
3	1.64	1.05	.28***	.0653	.34***	.0793	.20	--
4	1.60	.67	.26***	.0675	.13	.0807	.08	yes
Mean	--	--	*.38*		*.28*		--	
Median	--	--	*.36*		*.26*		--	
IPIP Variables								
Conscientiousness								
α = .85								
1	2.0	1.21	.57***	.0680	.63***	.0693	.72	--
2	1.87	1.31	.55***	.0685	.67***	.0690	.75	--
3	1.85	1.22	.64***	.0659	.28***	.0731	.49	no
4	1.80	1.19	.67***	.0650	.29***	.0725	.54	no
Agreeableness								
α = .68								
1	2.60	1.21	.54***	.0687	.18*	.0767	.33	no
2	1.81	1.25	.58***	.0681	.06	.0768	.34	no
3	2.26	1.43	.61***	.0678	.04	.0767	.38	no
4	2.98	1.16	.60***	.0679	.02	.0768	.36	no
Openness								
α = .75								
1	2.32	1.13	.44***	.0641	.32***	.0760	.29	yes
2	2.12	1.15	.81***	.0599	.22**	.0766	.71	no
3	2.32	1.19	.85***	.0597	.19**	.0769	.75	no
4	2.97	.91	.40***	.0658	.18**	.0771	.19	yes
Mean	--	--	*.61*		*.26*		--	
Median	--	--	*.59*		*.21*		--	
All Variables								
Mean	--	--	*.49*		*.27*		--	
Median	--	--	*.53*		*.23*		--	

*p<.05, **p<.01, ***p<.001

Table 3

Selected Fit Statistics and χ^2 *Values for Hierarchically Nested Factor Models*

Model	GFI	RMR	CFI	NNFI	RMSEA	χ^2	χ^2 df
1 (null)	--	--	--	--		1755.8	276
2	.8484	.0753	.8359	.8181	.0619	489.04	249
3	.6421	.1541	.3299	.2632	.1245	1231.4	251
4	.8019	.0996	.7343	.6770	.0840	615.7	227
5	.8986	.0646	.9203	.9022	.0454	341.7	225
6	.9018	.0587	.929	.9125	.0429	327.9	224
7	.8264	.0975	.7935	.7585	.0716	541.6	236
8	.8896	.0675	.9017	.8772	.0508	364.8	221

Table 4

χ^2Difference Tests between Hierarchically Nested Confirmatory Factor Models

Model Comparaison	χ^2 difference	χ^2 difference df	χ^2 crit.	p<	Issue addressed
6 vs. 3	903.5	27	55.476	.001	Convergent Validity
6 vs. 4	287.8	3	16.266	.001	Discriminant Validity
5 vs. 2	147.3	24	51.179	.001	Method Variance
6 vs. 5	13.8	1	10.28	.001	Covariation Between Method Factors

Table 5

Estimated Correlations Among Latent Variables in Model 6

Factors	*r*
Conscientiousness & Agreeableness	-.19
Conscientiousness & Openness	-.03
Agreeableness & Openness	.003
SJT Method & IPIP Method	.45

Table 6

Correlations Among Latent Trait Variables[1]

	Variable	1	2	3	4	5
1	sjtA	--	--	--	--	--
2	sjtC	.06	--	--	--	--
3	sjtO	-.40	.08	--	--	--
4	ipipA	*.61*	.32	.07	--	--
5	ipipC	.01	*.79*	.15	.27	--
6	ipipO	-.30	.06	*.86*	.08	.14

[1] Note: this analysis performed using a sample size of 394 available after the defense meeting.

Figures

Figure 1. Null model. Note that only two observed variables per trait for each method are shown for the purposes of simplicity. In actuality each trait was measured with 8 manifest variables (4 SJT and 4 IPIP).

Figure 2. Trait model. Note that only two observed variables per trait for each method are shown for the purposes of simplicity. In actuality each trait was measured with 8 manifest variables (4 SJT and 4 IPIP).

Figure 3. Method model. Note that only two observed variables per trait for each method are shown for the purposes of simplicity. In actuality each trait was measured with 8 manifest variables (4 SJT and 4 IPIP).

Figure 4. General trait model with method factors. Note that only two observed variables per trait for each method are shown for the purposes of simplicity. In actuality each trait was measured with 8 manifest variables (4 SJT and 4 IPIP).

Figure 5. Trait-method model with orthogonal methods. Note that only two observed variables per trait for each method are shown for the purposes of simplicity. In actuality each trait was measured with 8 manifest variables (4 SJT and 4 IPIP).

Figure 6. "Full" trait-method model with oblique methods. Note that only two observed variables per trait for each method are shown for the purposes of simplicity. In actuality each trait was measured with 8 manifest variables (4 SJT and 4 IPIP).

Figure 7. SJT method only model. Note that only two observed variables per trait for each method are shown for the purposes of simplicity. In actuality each trait was measured with 8 manifest variables (4 SJT and 4 IPIP).

Figure 8. SJT general trait model with method factors. Note that only two observed variables per trait for each method are shown for the purposes of simplicity. In actuality each trait was measured with 8 manifest variables (4 SJT and 4 IPIP).

Figure 1

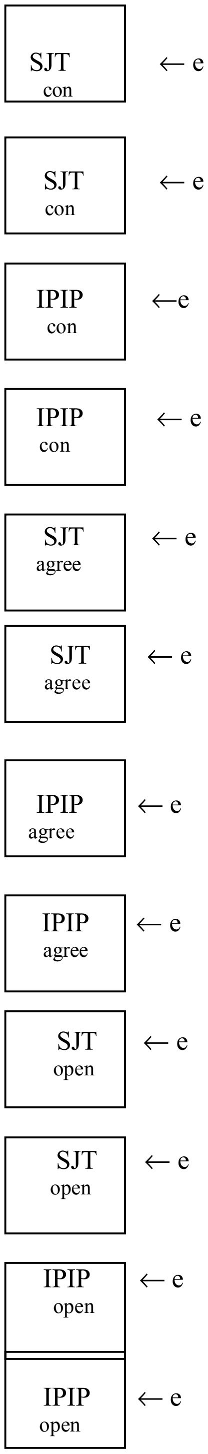

SJT con
← e
SJT con
← e
IPIP con
←e
IPIP con
← e
SJT agree
← e
SJT agree
← e
IPIP agree
← e
IPIP agree
← e
SJT open
← e
SJT open
← e
IPIP open
← e
IPIP open
← e

Figure 2

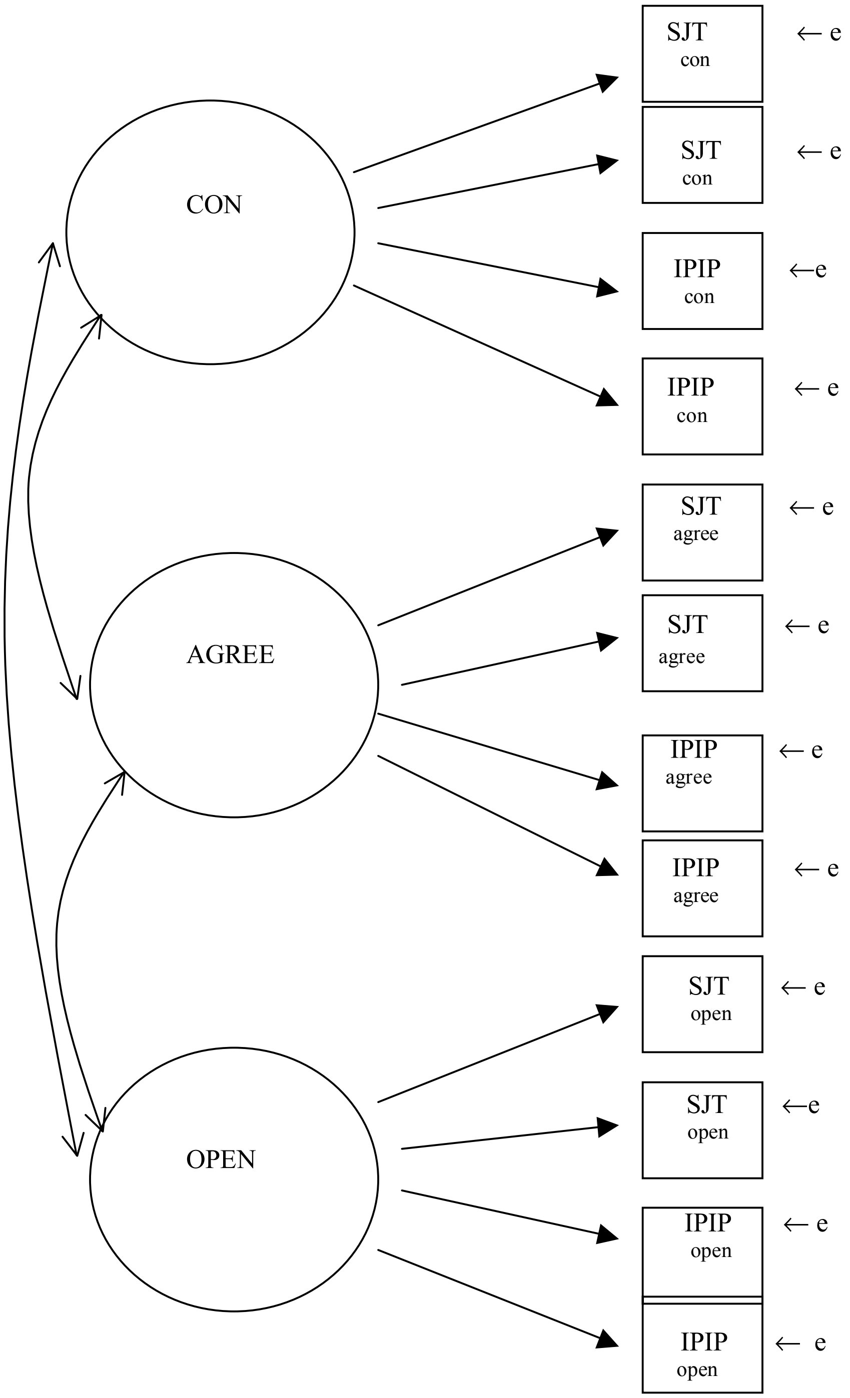

CON
AGREE
OPEN
SJT con
SJT con
IPIP con
IPIP con
SJT agree
SJT agree
IPIP agree
IPIP agree
SJT open
SJT open
IPIP open
IPIP open
← e
← e
←e
← e
← e
← e
← e
← e
← e
←e
← e
← e

Figure 3

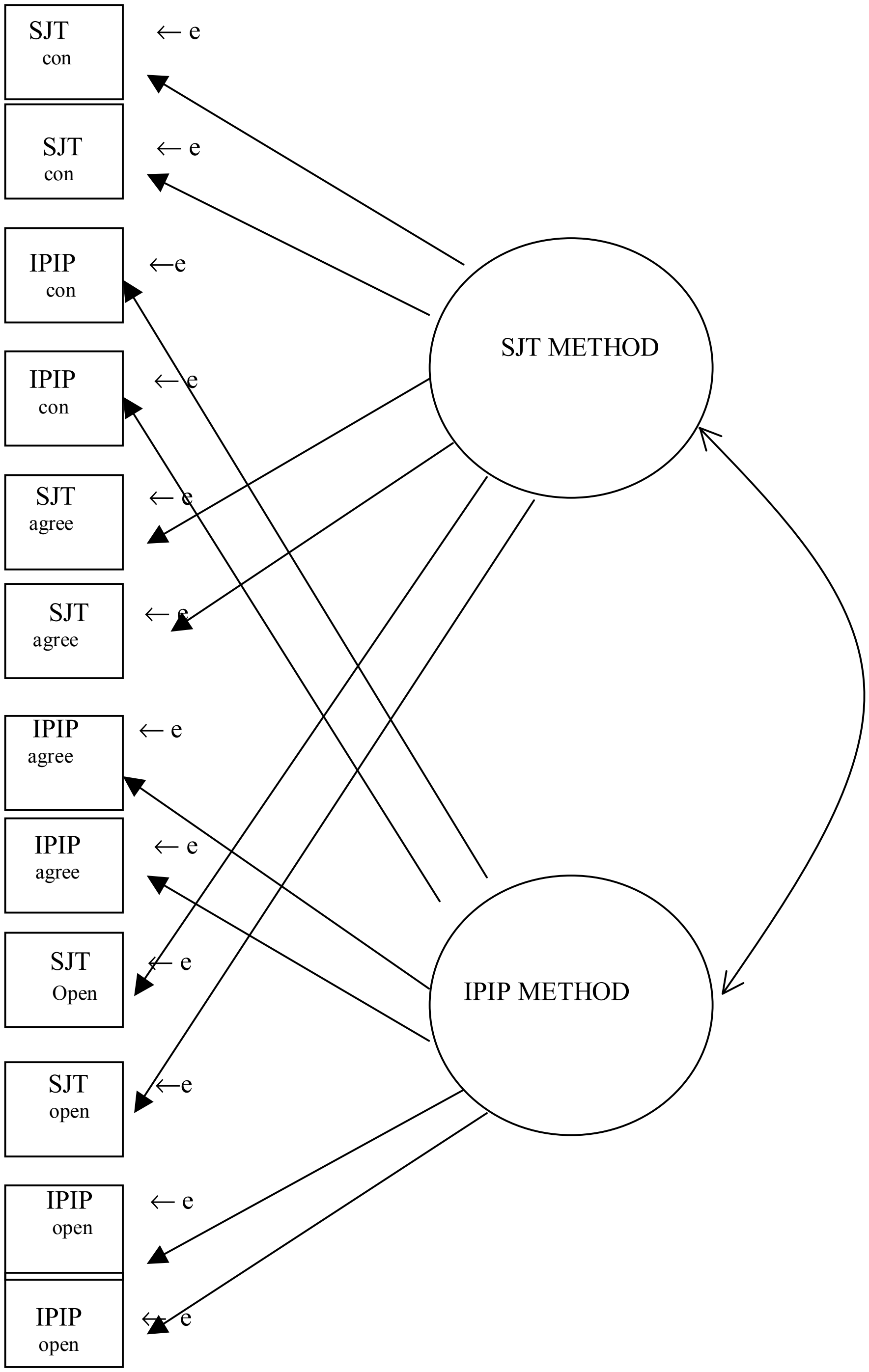

SJT con
← e
SJT con
← e
IPIP con
←e
IPIP con
← e
SJT agree
← e
SJT agree
← e
IPIP agree
← e
IPIP agree
← e
SJT Open
← e
SJT open
←e
IPIP open
← e
IPIP open
← e
SJT METHOD
IPIP METHOD

Figure 4

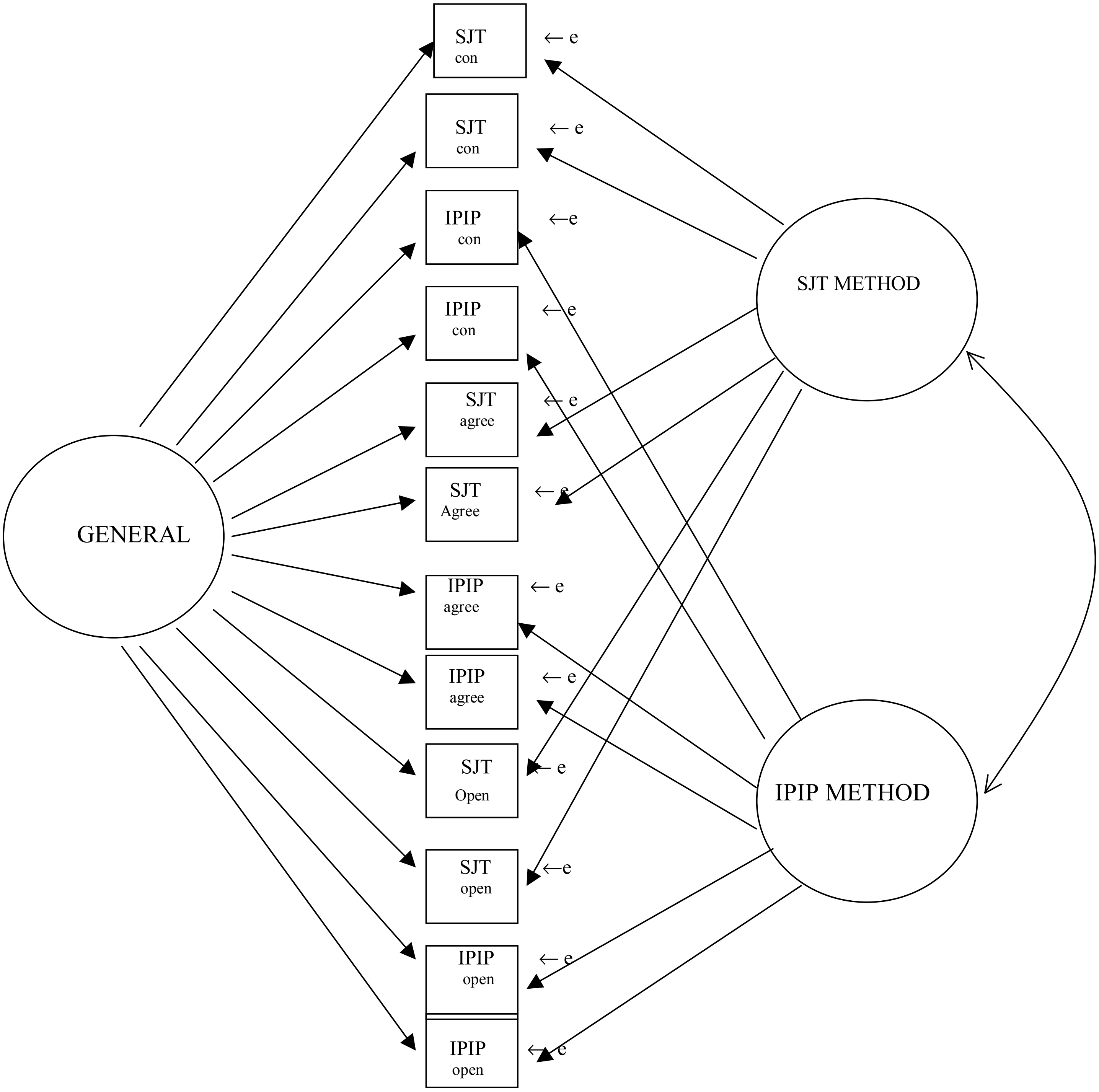

SJT con ← e
SJT con ← e
IPIP con ←e
IPIP con ← e
SJT agree ← e
SJT Agree ← e
IPIP agree ← e
IPIP agree ← e
SJT Open ← e
SJT open ←e
IPIP open ← e
IPIP open ← e
GENERAL
SJT METHOD
IPIP METHOD

Figure 5

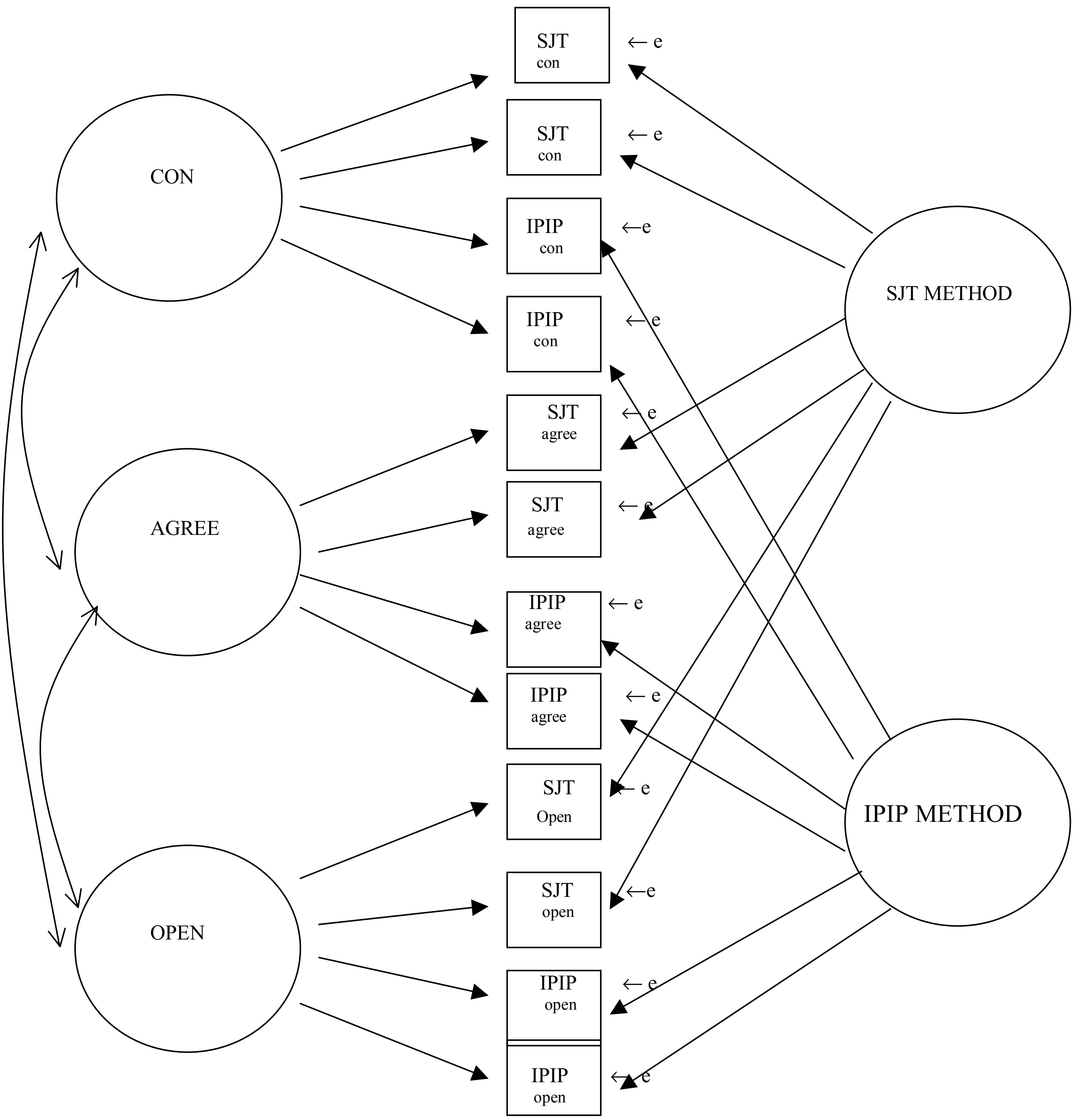

SJT con ← e
SJT con ← e
IPIP con ←e
IPIP con ←e
SJT agree ← e
SJT agree ← e
IPIP agree ← e
IPIP agree ← e
SJT Open ← e
SJT open ←e
IPIP open ← e
IPIP open ← e
CON
AGREE
OPEN
SJT METHOD
IPIP METHOD

Figure 6

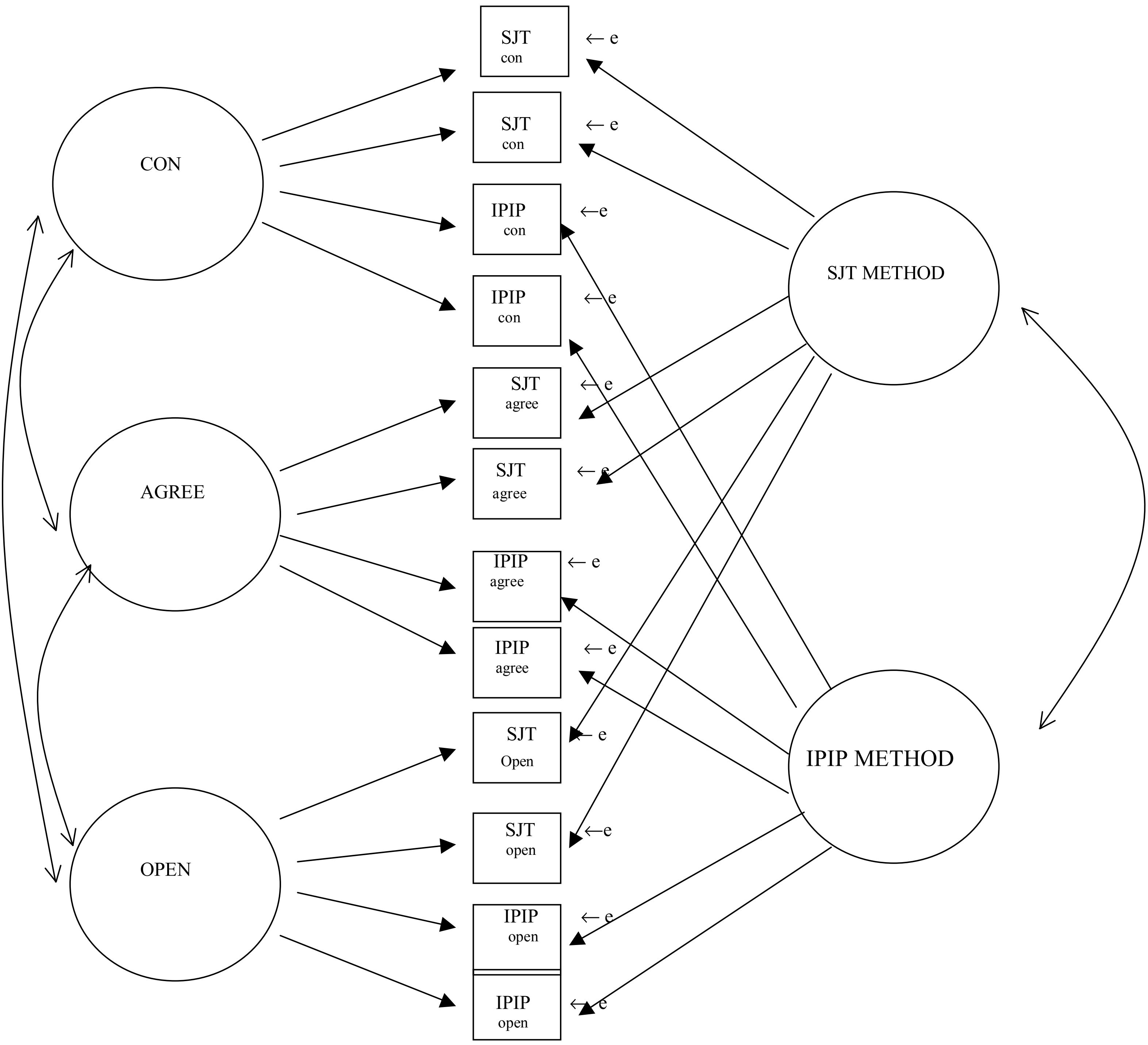

SJT con
← e
SJT con
← e
CON
IPIP con
←e
IPIP con
← e
SJT METHOD
SJT agree
← e
SJT agree
← e
AGREE
IPIP agree
← e
IPIP agree
← e
SJT Open
← e
IPIP METHOD
SJT open
←e
OPEN
IPIP open
← e
IPIP open
← e

Figure 7

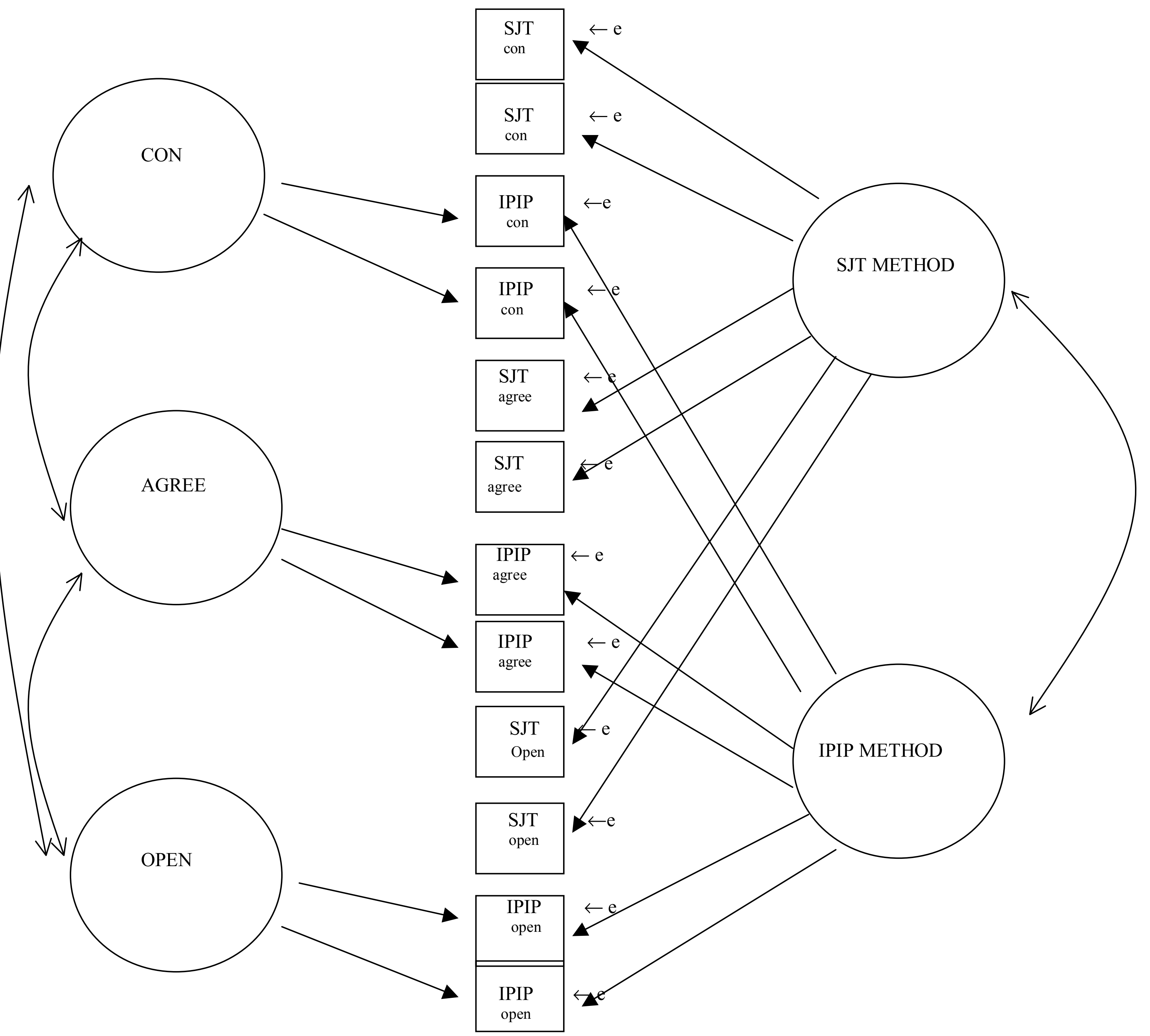
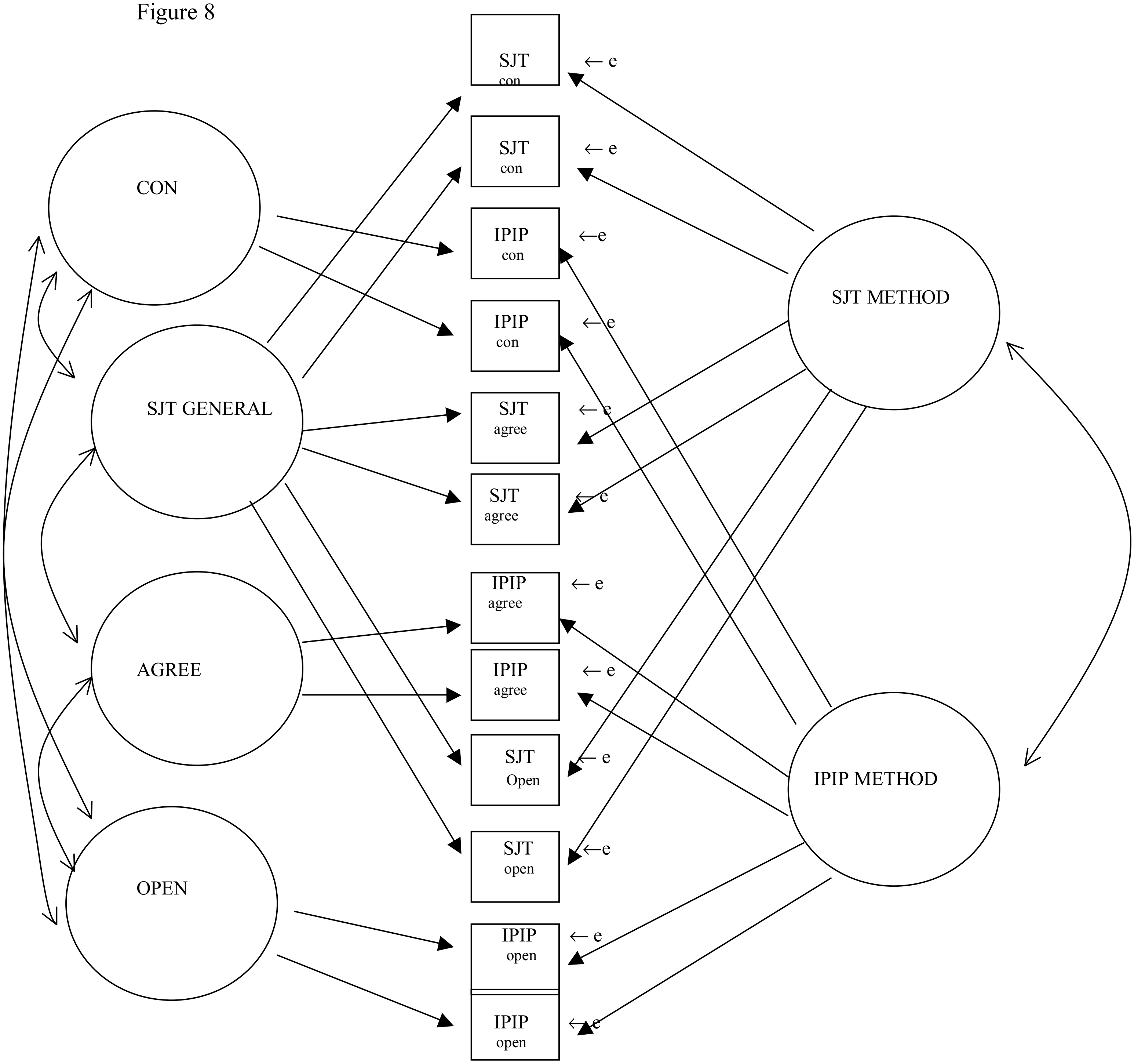

SJT con
← e
SJT con
← e
CON
IPIP con
←e
IPIP con
← e
SJT METHOD
SJT agree
← e
SJT agree
← e
AGREE
IPIP agree
← e
IPIP agree
← e
SJT Open
← e
IPIP METHOD
SJT open
←e
OPEN
IPIP open
← e
IPIP open
← e

Figure 8


SJT
con
← e
SJT
con
← e
IPIP
con
←e
IPIP
con
← e
SJT
agree
← e
SJT
agree
← e
IPIP
agree
← e
IPIP
agree
← e
SJT
Open
← e
SJT
open
←e
IPIP
open
← e
IPIP
open
← e
CON
SJT GENERAL
AGREE
OPEN
SJT METHOD
IPIP METHOD

Appendix A

IPIP Items

On the following pages, there are phrases describing people's behaviors. Please use the rating scale below to describe how accurately each statement describes *you*. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. So that you can describe yourself in an honest manner, your responses will be kept in absolute confidence. Please read each statement carefully, and then fill in the bubble that corresponds to the number on the scale.

Response Options

1: Very Inaccurate
2: Moderately Inaccurate
3: Neither Inaccurate nor Accurate
4: Moderately Accurate
5: Very Accurate

O4: ADVENTUROUSNESS

Prefer variety to routine.
Like to visit new places.
Interested in many things.
*Like to begin new things.
Prefer to stick with things that I know.
*Dislike changes.
*Don't like the idea of change.
Am a creature of habit.
Dislike new foods.
*Am attached to conventional ways.

A4: COOPERATION
Am easy to satisfy.
Can't stand confrontations.
Hate to seem pushy.
*Have a sharp tongue.
Contradict others.
*Love a good fight.
*Yell at people.
Insult people.
*Get back at others.
Hold a grudge.

C5: SELF-DISCIPLINE

*Get chores done right away.
Am always prepared.
*Start tasks right away.
Get to work at once.
Carry out my plans.
*Find it difficult to get down to work.
Waste my time.
Need a push to get started.
*Have difficulty starting tasks.
Postpone decisions.

Appendix B

INSTRUCTIONS:

For each of the following scenarios, please indicate your responses to the question that follows each in the blanks **ON THE OPSCAN FORM** you have been provided. When responding, please respond in terms of ***what you would actually do*** in each situation, not what you think the ideal or best response is. If a situation is unfamiliar or foreign to you, try to imagine yourself in that situation and answer according to how you think you would behave.

These items have been designed to measure aspects of personal preference—as such, ***there are no "right" or "wrong" answers***. Describe yourself as you generally are now, not as you wish to be in the future. So that you can describe yourself in an honest manner, ***your responses will be kept in absolute confidence***.

If you have a question, please raise your hand.

Please be sure to bubble in your student ID number—as this is how you will receive extra credit.

Please do not write on this form

Thank you for your participation

1. Class level:
(1) First Year
(2) Second Year
(3) Third Year
(4) Fourth Year
(5) Fifth or beyond

2. Ethnicity:
(1) African American
(2) Asian
(3) Caucasian
(4) Hispanic
(5) Other

3. Age:
(1) 18
(2) 19-20
(3) 21-22
(4) 23-25
(5) Over 25

4. Gender
(1) Female
(2) Male

*5	You and a number of others have been working as volunteers on an advertising campaign for a local charity for the last few months. You are becoming somewhat of an expert on this aspect of the project and are enjoying it quite a bit. The person in charge asks if a few volunteers could start working on a new phase of the project unrelated to the advertising campaign. *What would you do?* 1) Jump at the chance to begin work on the new phase of the project. 2) Try the new project for a little while. 3) Work on the new phase of the project only if no one else can. 4) Stick with aspect of the project you have been working on.
*6	It is the weekend and you have a large amount of work to do in the yard tomorrow. *What would you do?* 1) Sleep as late a possible and delay starting the yard work as much as you can. 2) Wakeup whenever and begin the yard work in the middle of the day. 3) Wake up whenever and begin the yard work soon after. 4) Set your alarm, get up early and begin working in the yard immediately.
*7	It is Thursday night. You have lots of homework to do and your parents are coming into town tomorrow afternoon. Your place is a mess and you need to do some grocery shopping. To complicate matters further, your friends are having a party tonight. *What would you do?* 1) Make a list of priorities—homework first, party last, cleaning and shopping in between--and carry out your plans in order of priority 2) Spend some time on each of your tasks and try to get everything done before you go to the party. 3) Study for a little while and go to the party—if you have time clean up in the morning. 4) Get the quick and easy things done and go to the party and worry about the rest later.
8	The supervisor is out of the office and an emergency arises that requires the immediate assignment of personnel. The secretary notifies you of the situation and explains that the supervisor is out of contact for at least three hours. The supervisor's immediate supervisor is also unavailable, but scheduled to return within the hour. *What would you do?* 1) Immediately make the assignment of needed personnel, and explain to the supervisor on his/her return. 2) Find another supervisor of appropriate status and advise them of the situation. 3) Simultaneously continue searching for one of the supervisors, and notify the personnel likely to be assigned to the emergency. 4) Call a meeting of the employees to discuss the situation. Collectively make a decision on what to do.
*9	It is the beginning of the semester and your professor notes that a significant portion of your grade will be based on a fairly comprehensive project that will be due the last day of class. *What would you do?* 1) Start working on the project right away. 2) Start working on the project at mid term. 3) Start working on the project near the end of the semester. 4) Start working on the project a few days before it is due.

***10**	You are about to graduate from college and are considering graduate school or jobs. You have a number of options with regard to location. *What would you do?* 1) Only consider schools and jobs that are similar in location to where you studied as an undergraduate. 2) Consider most schools and jobs that are similar in location and some that are new and different 3) Consider most schools and jobs that are in new and different places and some that are similar in location. 4) Only consider schools and jobs that are in new and different locations
11	You are an experienced employee. A new employee comes to you for assistance. You spend time showing the employee how to do a task. Next month the same thing happens and you again help the new employee do the same task. This situation continues and you finally get upset since the new employee should be able to do the task alone. *What would you do?* 1) Explain to the person that you do not understand what the problem is with the task but that you have helped as much as you can. 2) As long as the person was trying, continue to show the person how to do the task. 3) Ask the employee to take notes or make a copy of the product to use as a guide in the future for how to perform the task. 4) Inform the employee to pay careful attention because this is the last time you will demonstrate how to do the task. 5) Sit down with the employee to try to determine what the problem is so that you can figure out the best way to deal with the situation from here on.
12	You are assigned a complicated project, and collect hundreds of boxes of documents that need to be inventoried, reviewed, and evaluated. You ask your supervisor for the help of a computer person, but there are none available for a month. You can't operate a computer, but you know this is a priority matter. *What would you do?* 1) Ask the supervisor to find the available help from another source. 2) Seek funds to hire a temp. 3) Request training to learn the computer skills that you need. 4) Solicit help from co-workers more computer literate than yourself. 5) Have a clerk start the inventory manually until a person with computer skills becomes available.
***13**	You are at an art show and stop for a few seconds to look at a painting. You think it is truly awful. You think to yourself that whoever painted this particular piece is completely lacking in talent. Just then, a stranger next to you comments on how brilliant the painting is and says the artist who painted it is a genius. *What would you do?* 1) Agree with the stranger that the painting is very nice. 2) Nod and say that that the painting is ok. 3) Say that the painting is ok, but that you have seen much better 4) Contradict what the stranger has said, adding your opinion that the painting is awful

*14	You are in class and the student next to you makes a comment to the class that you know to be factually inaccurate. *What would you do?* 1) Tell the student what he said is wrong. 2) Raise your hand and ask the professor if what the student said was correct. 3) Talk to the student after class about what he said. 4) Say nothing.
15	You are in the office and get a telephone call that requires immediate attention. To perform effectively, you need to change the conditions of a contract, which technically needs approval. Your supervisor is out and cannot be reached. If you change the conditions of the contract, you may be reprimanded later for not having proper approval. If you do nothing, the contract may be lost. *What would you do?* 1) Seek approval from the next person in the chain of command. 2) Make a decision about which is more important to the company, the technical requirement or ensuring that the contract is maintained, and act accordingly. 3) Do what it takes to close the contract and follow up with documentation to justify your actions. 4) Follow the procedures and risk the contract if necessary.
16	You are involved in a reading intensive class with 15 other students. As a check to ensure everyone is doing the readings, your professor randomly draws a students name out of a hat every class. The student whose name is drawn must present a short summary of the readings due that day to the rest of the class. *What would you do?* 1) Prepare a short summary for every class. 2) Prepare a short summary for most classes. 3) Prepare a short summary every once in awhile. 4) Prepare nothing and hope you are not called on.
*17	You are planning a vacation you will be taking in a few months. The travel industry is not very strong, so almost all vacation packages are an excellent value. You are trying to decide where to go among a number of packages that cost about the same. *What would you do?* 1) Try a new vacation spot in a foreign location and different culture. 2) Try a new vacation spot in a familiar location and culture. 3) Vacation in a place you have been once before and enjoyed. 4) Vacation in a place you have been a number of times and know is nice.
18	You are riding the bus and the person next to you is telling you and the people around you about a recent local event that is completely unbelievable and a clearly a distortion of the truth. *What would you do?* 1) Smile and nod in agreement with the person, and say you agree with the story 2) Smile and nod in agreement with the person, but say nothing 3) Nod in agreement with the person, but say nothing 4) Nod in agreement with the person, then correct his or her story

***19**	You are walking through campus and pass a group of students holding a demonstration. One of the demonstrators yells at you for no reason. *What would you do?* 1) Keep walking without saying a word. 2) Quietly ask the demonstrator not to yell at you. 3) Yell back at the demonstrator, telling him/her to be quiet. 4) Yell back at the demonstrator and insult him or her.
20	You are working on a group project for class, and the project is due in a week. The assignment is to write a research paper on the material you've been working on with your group since the beginning of the semester. One person takes charge as leader for the group, and deals out what each person should work on. It is clear that you wound up with the bulk of the work, and everyone else seemed to have gotten off with an easier workload. *What would you do?* 1) Refuse to do the work you have been chosen to do and assert to the group that the workload is unevenly distributed. 2) Point out to your group that it seems the workload is a bit unfair, and try to see if they can change their minds about who does what. 3) Accept the assignment, but act noticeably unhappy about it 4) Take the assignment and leave without saying a word, better to just go ahead and do it rather than cause a fuss.
***21**	You go to a new restaurant in town. Some of what is on the menu is unfamiliar to you and somewhat exotic. *What would you do?* 1) Order something new and exotic that you have never tried before. 2) Order something you have never tried but have heard of. 3) Order something that sounds like a variation of something familiar to you. 4) Order something very familiar to you.
***22**	You have a group project due in one of your classes. Your group agreed on Monday to start working on the project over the weekend. You have not heard from them since then. *What would you do?* 1) Wait until someone from your group contacts you to give you the push you need to start working on the project. 2) Wait until it is later in the weekend to give you the push you need to start working on the project. 3) Start working on the project later in the weekend without really needing a push to get started. 4) Start working on the project early in the weekend without really needing a push to get started.
23	You have just moved to a new community. You receive a community newsletter in the mail that has a long list of all the activities and events that regularly occur in the community. You can sign up to participate in as many or as few as you would like. *What would you do?* 1) Get interested in many different events and activities. 2) Get interested in a few events and activities. 3) Get interested in one event or activity. 4) Decide you are not interested in much.

24	You live about ten miles away from campus and drive to school everyday. There are number of different ways you can drive that all take about the same amount of time. *What would you do?* 1) Drive the same way to school every day 2) Drive the same way to school most days 3) Drive a different way to school most days. 4) Drive a different way every day.
25	You observe another employee performing her job in a questionable manner. The other employee is assigned to a different group, but her performance could affect the outcome of your project. On the other hand, if you report the employee, it could cause friction with the other group and that could affect the long term working relationship between the two groups. *What would you do?* 1) Accept the employee's deficiencies and work around them. 2) Discuss your problems with the employee, and inform her that if her performance does not improve you will be forced to report her. 3) Without making the problem employee suspicious, get another worker to aid her in getting her work done. 4) Advise a supervisor of the problem. 5) Set up a group meeting in which all employees discuss the problems they have encountered in achieving the group's goals. Bring up this employee if necessary.
26	You return from spring break and realize you have a number of bills to pay, a pile of laundry to do, as well as some general cleaning to do around your place. *What would you do?* 1) Get your chores done right away. 2) Get your chores done within the day. 3) Get your chores done in a day or two. 4) Leave the chores undone.
***27**	Your class is involved in a class discussion and your instructor makes a comment that you strongly disagree with. *What would you do?* 1) Respectfully confront your instructor right then and there. 2) Respectfully confront your instructor after class about his/her comment. 3) Do not confront your instructor. 4) Nod in agreement with your instructor.
28	You're working on a deal that involves a lot of paperwork and record reviewing with another employee. You take leave and come back to find that your desk is filled with more records. The employee you were working with re-arranged your papers and desk, and you can't tell what was already reviewed and what you've done. You're angry. *What would you do?* 1) Take a few minutes to cool down and then request that the employee report what progress has been made in written and oral form. 2) Ask the co-worker if he/she knew where you had left off since the records are now shifted. 3) Inform the employee that you are angry for what he/she has done and tell the employee that in the future you do not want him/her to touch your desk. 4) Assume that the employee has no bad intentions, and try to re-sort the piles. 5) Realize it is your fault for leaving the files unattended on your desk.

D. Matthew Trippe

2300 San Marcos Street, Blacksburg, VA 24060, 540-552-7542, (dtrippe@vt.edu)

Education

2002-Current Virginia Tech Blacksburg, VA

***Ph.D. Industrial-Organizational Psychology**

- *Expected Graduation: Spring, 2004

2000-2002 Virginia Tech Blacksburg, VA

M.S./Industrial-Organizational Psychology

- GPA: 3.9
- Graduate Assistant: Institutional Research Planning & Analysis
- Graduate Teaching Assistant
  - Instructor: Introduction to Psychology Recitation
  - Instructor: Laboratory in Advanced Social Psychology

1995-1999 College of Charleston Charleston, SC

B.S./Psychology, Minor: Philosophy

- GPA: Overall 3.4, Major 3.5

Work Experience

1999-2000 King & Spalding Atlanta, GA

Document Clerk: Toxic Tort Litigation Team

- Create and maintain intricate document database for medical records and production documents in large scale/multi-million dollar litigation.
- General document support for attorneys.

Summer 1998 Sperduto & Associates Atlanta, GA

Intern: I/O Consulting Firm

- Search scholarly literature for project specific information
- Provide aggregate/summary reports of psychometric data

Summer/Christmas 1997 Carmax Atlanta, GA

Sales Associate

- Sale of new and used vehicles
- Responsible for finance aspect of each sale
- Product presentations to sales team

Professional Memberships

SIOP: Society for Industrial Organizational Psychology (student affiliate)

APA: American Psychological Association (student affiliate)

Psi Chi: The National Honor Society in Psychology

Computer Skills

Microsoft Excel, Work, PowerPoint, Access, Outlook.
SAS, SPSS, MULTILOG, BILOG, ANALOG