Interrater Agreement of Incumbent Job Specification Importance Ratings: Rater, Occupation, and Item Effects

Steven R. Burnkrant

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

Doctor of Philosophy

in

Industrial and Organizational Psychology

Neil M. A. Hauenstein, Chair
Robert J. Harvey
Roseanne J. Foti
Robert S. Stephens
Kevin D. Carlson

October 23, 2003
Blacksburg, Virginia

Keywords:
Job specification, Job analysis, KSAO, Competency, Interrater agreement, Validity

Interrater Agreement of Incumbent Job Specification Importance Ratings: Rater, Occupation, and Item Effects

Steven R. Burnkrant

Abstract

Despite the importance of job specifications to much of industrial and organizational psychology, little is known of their reliability or validity. Because job specifications are developed based on input from subject matter experts, interrater agreement is a necessary condition for their validity. The purpose of the present research is to examine the validity of job specifications by assessing the level of agreement in ratings and the effects of occupational tenure, occupational complexity, and the abstractness of rated worker requirements. Based on the existing literature, it was hypothesized that (1) agreement will be worse than acceptable levels, (2) agreement will be higher among those with longer tenure, (3) agreement will be lower in more complex occupations, (4) the effect of occupational tenure will be more pronounced in complex than simple occupations, (5) agreement will be higher on more abstract items, and (6) agreement will be lowest for concrete KSAOs in complex occupations. These hypotheses were tested using ratings from 38,041 incumbents in 61 diverse occupations in the Federal government. Consistent with Hypothesis 1, agreement failed to reach acceptable levels in nearly every case, whether measured with the a_{wg} or various forms of the r_{wg} agreement indices. However, tenure, occupational complexity, and item abstractness had little effect on ratings, whether agreement was measured with r_{wg} or a_{wg}. The most likely explanation for these null findings is that the disagreement reflected a coarse classification system that overshadowed the effects of tenure, complexity, and abstractness. The existence of meaningful subgroups within a single title threatens the content validity of job specifications: the extent to which they include all relevant and predictive KSAOs. Future research must focus on the existence of such subgroups, their consequences, and ways of identifying them.

Dedication

I dedicate my dissertation to all those who supported me, especially my parents, Richard and Janis, who never doubted or dissuaded, and my loving wife, Jeanne, who stood steadfastly by my side through it all.

Acknowledgements

I would like to thank all the teachers in my life, whose guidance and wisdom have brought me this far. First, I credit the outstanding education I have received, at Sierra Elementary, Oberon Junior High, Arvada West High School, and particularly Colorado College and, of course, Virginia Tech. Second, I credit my committee, especially Neil Hauenstein, for shaping my thinking, suggesting improvements to my research design, and always encouraging me to finish. Third, I would like to thank the U.S. Office of Personnel Management for making available the datasets used in this dissertation. Finally, and most notably, I would like to thank the countless people, especially my wife and family and friends, who taught me life’s little lessons, including the most important: You can do it.

Table of Contents

List of Tables

List of Figures

Interrater Agreement of Incumbent Job Specification Importance Ratings: Rater, Occupation, and Item Effects

INTRODUCTION

Job specifications, or the knowledge, skills, abilities, and other characteristics (KSAOs) required by an occupation, form the basis of such personnel practices as recruitment, selection, performance appraisal, training, and vocational guidance. As such, they lie at the heart of much of industrial and organizational psychology. Despite their central role, there is a paucity of research on the reliability and validity of KSAO ratings. Very little is known about such fundamental issues as the degree to which subject matter experts (SMEs) agree in their ratings and the moderating roles of incumbent, occupational, and methodological factors. Unfortunately, the validity of KSAO ratings is taken, without justification, as almost axiomatic.

The validity of KSAO ratings is by no means a trivial issue and cannot be taken for granted. At best, an invalid job specification will result in wasted resources, a consequence of, for example, inappropriate selection systems or training programs; at worst, an invalid job specification will result in incompetent incumbents in critical positions that should have gone to more qualified applicants. These concerns have been codified in both legal (e.g., *Contreras v. City of Los Angeles*, 1981) and professional (e.g., *Principles for the Validation and Use of Selection Procedures*, 1987) requirements, and underscore the need for additional research on the job specification process.

Because job specifications describe group-level requirements, they are valid only to the extent that the requirements apply to all incumbents in the occupation. Interrater agreement among SMEs, therefore, is a necessary but not sufficient requirement for the validity of KSAO ratings (Harvey, 1991). The limited research that has been conducted on KSAO ratings (e.g., Jones et al., 2001), in conjunction with research on such related topics as job analysis (e.g., Dierdorff & Wilson, 2003), job evaluation (e.g., Hahn & Dipboye, 1988) and performance appraisal (e.g., Conway & Huffcutt, 1997), suggests that KSAO ratings may often be invalid because interrater agreement frequently falls below acceptable levels. Such findings are problematic given the recent trend toward more abstract methods of setting worker requirements, such as "competency modeling," that rely on less rigorous methods (Schippmann et al., 2002) and that may not use SMEs at all (Harris, 1998).

Therefore, the first purpose of the present research was to examine interrater agreement using a large sample of diverse occupations. The variety of occupations allows a much broader and more generalizable assessment of agreement levels than is currently available in the literature. The second purpose was to explore the degree to which occupational tenure, occupational complexity, and KSAO abstractness influence interrater agreement. Identifying the effects of these factors will help practitioners understand the sources of variability in KSAO ratings and potentially identify sources of variability that reduce the validity of KSAO ratings.

The first section of the literature review is devoted to a discussion of background information, which is designed to, first, introduce the terminology used throughout this paper, and, second, emphasize the importance of interrater agreement in terms of both the methods used to construct job specifications and the legal and professional standards that govern their development and use. Different measures of agreement are reviewed next, to establish a suitable context in which to discuss prior research. After reviewing the extant literature to show that acceptable levels of agreement are seldom reported, sources of disagreement are reviewed, with particular attention paid to occupational tenure, occupational complexity, and KSAO abstractness. After presenting the hypotheses, the method and analyses are described.

LITERATURE REVIEW

Background

Terminology

To avoid as much as possible the semantic confusion in job analysis terminology noted by Harvey (1991), this section will define the terms used throughout this paper. First, job characteristics can be analyzed at three main levels, the element, the task, and the duty (Cascio, 1998). An *element* is the smallest unit of analysis, describing a highly specific action carried out for a highly specific purpose (e.g., pressing keys on a typewriter to mark letters on a piece of paper). Comprised of more than one element, a *task* is a more general activity performed for a more general purpose (e.g., typing a letter). The broadest job characteristic is a *duty*, which comprises any number of tasks collectively carried out for a general purpose (e.g., communicating with the public).

Second, worker characteristics describe the knowledge, skills, abilities, and other characteristics (KSAOs) that employees must possess to perform a job's elements, tasks, and duties. In this context, *knowledge* is having the necessary information to perform one's job (e.g., fluency in English) and a *skill* is the competence to perform a learned activity (e.g., typing). *Abilities* and *other characteristics* describe unobservable worker characteristics, such as intelligence (in the case of ability) and personality traits (in the case of other characteristics) that may be correlated with the KSs required by a job and the ability to successfully carry out the relevant elements, tasks, and duties.

Third, workers may be classified by their position, job, occupation, and occupational family. A *position* describes the work of an individual employee. A *job* is a group of similar positions within a single organization. The same or similar jobs across multiple organizations are collectively referred to as an *occupation*. At the highest level, an *occupational family* refers to groups of similar occupations (e.g., clerical, professional). In the Federal government, workers are classified according to occupational *series*, where each series refers to similar jobs in different agencies. Within series, workers are classified by *grade*, a classification that corresponds to the level of responsibility, job complexity, and other characteristics of a position in a series, as well as by *step*, a sub-classification of grade primarily determined, in practice, by organizational tenure rather than any work requirements or responsibilities. The private sector equivalents of grades and steps are usually more fluid and less rigidly defined, as are the accompanying compensation systems.

Finally, given job and worker characteristics on the one hand, and jobs and occupations on the other, one can distinguish among job analyses, job specifications, occupational analyses, and occupational specifications. Following Harvey (1991; see also Harvey & Wilson, 2000), *job analysis* is "the collection of data describing (a) observable (or otherwise verifiable) job behaviors performed by workers, including both *what is accomplished* as well as *what technologies are employed* to accomplish the end result and (b) verifiable characteristics of the job environment with which workers interact, including physical, mechanical, social, and informational elements" (p. 74, emphasis in original). By extension, an *occupational analysis* refers to the description of observable job behaviors and characteristics common to the jobs classified in a single occupation. A *job specification*, on the other hand, refers to a description of the worker characteristics needed to perform a job successfully (i.e., the setting of KSAO requirements). One may then define an *occupational specification* as a description of worker characteristics needed to perform the requirements of an occupation. A worker with the KSAOs identified by an occupational specification should be able to transfer with minimal difficulty to any job within that occupation.

Methods of Setting KSAO Requirements

The various methods of developing a job specification all rely on subject matter experts (SMEs) to rate the applicability and importance of KSAOs. The SMEs may be incumbents, supervisors, or job analysts; ratings may be made directly or with reference to a job analysis; and ratings may be made using a questionnaire or through discussion by a panel of SMEs. With *threshold traits analysis* (Lopez, Kesselman, & Lopez, 1981), for example, supervisors in a group first independently and directly rate the relevance and level required of 33 traits, and then resolve disagreements with a discussion. Using an adjective checklist approach (Arneson & Peterson, 1986, cited in Harvey, 1991), trait specifications are inferred from incumbents' direct ratings of the relevance of various worker characteristics. In other cases (e.g., Hughes & Prien, 1989), a job analyst may conduct a "pilot" job analysis (e.g., via observation) to identify relevant tasks, behaviors, and KSAOs, and from which a standard questionnaire will be developed. SMEs then rate the job tasks and KSAOs on the same questionnaire. Those KSAOs meeting some specified cutoff are then used, for example, in a selection system. To save time and money, some have advocated the use of direct, holistic ratings of worker requirements, an approach adopted by some "competency modelers" (e.g., Shippmann et al., 2002) and the *Occupational Information Network* (O*Net). Whatever the method, individual ratings must be aggregated to form summary scores for the job or occupation.

The question then becomes whether or not the resulting job specification is valid. As discussed by Harvey (1991), if KSs are set with reference to a task- or behavior-based job analysis, an analyst may easily verify their necessity by matching job requirements with the skills and knowledge needed to perform them (i.e., content validation). On the other hand, when

KSs are rated directly, without reference to a job analysis, and whenever AOs are rated, validity can only be inferred. Such evidence may come from a criterion-related validation study or from the use of job component validation (e.g., McCormick, DeNisi, & Shaw, 1979). Even these methods, however, presuppose the existence of internally valid KSAO ratings. In other words, whether SMEs rate KSAOs directly or with reference to a job analysis, adequate interrater agreement is a necessary but not sufficient precondition for the validity of the resulting job specification.

Legal Requirements and Professional Guidelines

While not explicating how one should conduct a job specification, legislation and its accompanying case law clearly require that they be valid and job-related. Title VII of the *Civil Rights Act of 1964* (and its extensions: the *Equal Employment Opportunity Act of 1972* and the *Civil Rights Act of 1991*), the *Equal Pay Act of 1963*, the *Age Discrimination in Employment Act of 1967*, and the *American with Disabilities Act of 1990* each has as its central objective the prevention of employment practices that deny equal opportunity to otherwise qualified members of protected groups. Because they delineate occupational requirements, comprehensive and thorough job analyses form the backbone of legally defensible personnel practices (*Albemarle Paper Co. v. Moody*, 1975; *EEOC v. Atlas Paper Box Co.*, 1989; *Griggs v. Duke Power Co.*, 1971; *Jones v. New York City Human Resources Administration*, 1975; *Sledge v. J. P. Stevens & Co.*, 1978). In *Guardians Association of the New York City Police Department v. Civil Service Commission of the City of New York* (1980), the court ruled as impermissible a test constructed to measure KSAOs identified by a perusal of occupational materials (e.g., job manuals), without conducting a formal job analysis (e.g., observing worker behavior). Along the same lines, the court ruled in *U.S. v. State of New York* (1979, cited in Harvey, 1991) that direct ratings of

KSAO criticality, in the absence of a task-based job analysis that demonstrates their relevance to the job, are not sufficient to justify their inclusion in an employment test (cf. Kesselman & Lopez, 1979). Echoing these and other decisions (e.g., *Vulcan Society v. Civil Service Commission*, 1973), the court in *Contreras v. City of Los Angeles* (1981) ruled that KSAO requirements must be shown to be predictive of important work behavior. Importantly, the court viewed the agreement in KSAO criticality ratings as evidence that the ratings were neither skewed nor biased. In other words, a KSAO should not be included in an employment test if job experts do not agree on its importance. Thus, the KSAOs used in personnel practices should be linked to verifiable occupational requirements and must be clearly critical, as judged, for example, from agreement in ratings of importance.

Professional guidelines, viz. the *Uniform Guidelines on Employee Selection Procedures* (1978), the *Standards for Educational and Psychological Testing* (1985), and the *Principles for the Validation and Use of Selection Procedures* (1987), generally parallel the requirements set forth by legislation and case law. The *Guidelines*, for example, have a clear preference for minimizing the inferential leap from job content to required KSAOs, stating "a selection procedure based upon inferences about mental processes cannot be supported solely or primarily on the basis of content validity" (section 14C(1)). The *Standards*, though not specific with respect to job analysis, similarly note that content validation requires a direct link between a test and job content, and that construct validation must be based on more than expert judgment alone. The *Principles* are somewhat less strict and somewhat more vague, as in the requirements that "scales used to evaluate tasks and [KSAOs] should have *reasonable* psychometric properties" and "lack of consensus should be *noted* and carefully *considered*" (p. 5, emphasis added). The meaning and implications of these terms are not exactly clear, although it can be inferred from the *Principles'* recommendation that construct validation be used for general KSAOs that KSAO ratings should be reliable and valid. When multiple raters are used, this implies that, at a minimum, SMEs should agree on the relevance and importance of the KSAOs (Harvey, 1991).

Measures of Agreement

A great number agreement measures have been proposed, ranging from simple percentage agreement, interrater correlations, and the standard deviation (Schmidt & Hunter, 1989) to more sophisticated models such as the Content Validity Ratio (Lawshe, 1975), the *T* index (Tinsley & Weiss, 1975), Kappa (Cohen, 1960) and its extensions (Berry & Mielke, 1988; Janson & Olsson, 2001), Maxwell's random error coefficient (Maxwell, 1977), a latent-class model (Schuster & Smith, 2002), the average deviation index (Burke & Dunlap, 2002; Burke, Finkelstein, & Dusig, 1999), intraclass correlations (Shrout & Fleiss, 1979), within-and-between analysis (Dansereau, Alutto, & Yammarino, 1984), generalizability theory (Chronbach, Gleser, Nanda, & Rajaratnam, 1972), the r_{wg} class of indices (e.g., James, Demaree, & Wolf, 1984; Lindell, Brandt, & Whitney, 1999), and, most recently, the a_{wg} index (Brown & Hauenstein, 2003). In applications in which ratings are made on a continuum (i.e., agreement is not a simple all-or-none dichotomy), the historically most popular measures of agreement have been the average of all possible interrater correlations and, secondarily, the intraclass correlation. As is described below, however, these measures have serious limitations and are generally inferior to generalizability theory and the r_{wg} and a_{wg} indices for use in assessing agreement of KSAO ratings.

The majority of studies that examine interrater agreement use the average of all possible interrater correlations (IRCs), a measure that is typically inappropriate and often misinterpreted. First, IRCs are often referred to as indexing interrater agreement, although IRCs are correlational in nature and measure profile similarity, not the extent to which different raters give interchangeable ratings (Kozlowski & Hattrup, 1992). As such, IRCs require multiple objects of measurement and cannot be used to judge agreement on a single item. Second, IRCs cannot be computed when raters agree perfectly. Third, although IRCs are most often used to measure interrater reliability, they can properly be interpreted as such only when the different raters function as parallel measures (Murphy & DeShon, 2000a), an assumption seldom tested or explicated (e.g., Schmidt, Viswesvaran, & Ones, 2000). Fourth, some authors have used IRCs to draw inferences about test validity. Murphy and DeShon, (2000b; cf. Fleishman & Mumford, 1991), for example, approaching IRCs from a classical test theory perspective, note that true scores represent the expected value of the observed ratings, not the rated construct itself, yet describe measurement error as indexing sources of invalidity in observations—a logically untenable position given that validity concerns the degree to which the true score represents the rated construct, not the extent to which observed scores represent the true score. For these reasons, IRCs should simply be interpreted as average profile similarity, not agreement, reliability, or validity.

Intraclass correlations (ICCs) measure the proportion of variance in ratings due to the objects of measurement (McGraw & Wong, 1996; Shrout & Fleiss, 1979). For example, assuming multiple SMEs rate several KSAOs, a high ICC indicates that the variance in the ratings is due to the different KSAOs, not the different raters. Of the many varieties of ICCs, the two general models most relevant here are ICCs for consistency and those for agreement. Unlike ICC(C)s, ICC(A)s include within-row between-column (e.g., within-rater between-KSAO) variance as a term in the denominator. That is, ICC(A)s take into account exact agreement, not just relative agreement in terms of profile similarity. For this reason, ICC(A) is the more appropriate model for assessing convergence in job analysis and KSAO ratings. However, ICCs have serious drawbacks, the most apparent of which is the requirement of multiple targets: ICCs cannot be used to measure convergence on a single object of measurement. One could not, for example, examine agreement on ratings of a single KSAO. Second, and related, ICCs give only a single summary statistic for convergence on all targets collectively. Third, ICCs cannot be calculated when raters agree perfectly. Finally, like all methods derived from the Analysis of Variance framework, ICCs are based on the assumption of equal column (e.g., KSAO) variances. ICCs, therefore, offer limited utility as measures of interrater agreement.

Generalizability theory (G-theory; Cronbach et al., 1972; Shavelson & Webb, 1991) represents a significant advance beyond ICCs and especially IRCs (e.g., Murphy & DeShon, 2000a). Unlike the classical test theory perspective, which is embodied by IRCs and ICCs, G-theory explicitly recognizes that variance is composed of more than orthogonal true scores and random error. Rather, in most cases, variance is in part a function of rater characteristics, item characteristics, and other factors that are neither true score nor random error. Once these sources of variance are taken into account, one may estimate the degree to which the observed scores are likely to generalize to the universe of possible scores. In effect, G-theory estimates the appropriateness of generalizing one rater’s score to another rater’s score.

A G-theory analysis typically proceeds in two steps. In the first step, called the generalizability study, variance components are estimated from an ANOVA model, using, for example, estimated Mean Squares obtained from an ANOVA and the formulas of Cardinet, Tourneur, and Allal (1976), Minimum Norm Quadratic estimation, or Restricted Maximum Likelihood estimation (see Searle, Casella, & McCulloch, 1992). Variance components, and the proportion of variance they contribute to observed ratings, can be computed for each main effect and each interaction, except for the highest-order interaction which, because each cell in the model contains a single observation, is confounded with an undifferentiated error term.

In the second step, called the decision study, a G-coefficient is obtained that represents the ratio of the universe (true) score variance of the facet of interest (i.e., the facet one wishes to generalize) to the observed score variance—the larger the ratio, the more dependable the measurement. For example, as in the present study, if one wishes to generalize an item rating, then the variance component for item is divided by the sum of the variance components for the facet of interest and the relevant sources of error (e.g., raters). If the researcher is concerned only with relative standing (e.g., the rank order of KSAO importance), then the error term comprises the interactions of each facet with the facet of interest, and the resulting coefficient, ρ^2, is an intraclass correlation. If, on the other hand, the researcher wishes to make an absolute decision (e.g., the level of KSAO importance matters), then the facet main effects also contribute to error, and the resulting coefficient, Φ, is called the index of dependability (Brennan & Krane, 1977). Because, in each of these models, the variance components for the facets that make up the error are divided by the number categories in that facet, one may estimate ρ^2 and Φ under different conditions (e.g., 5 raters vs. 10 vs. 20): the greater the number of categories, the smaller the error term, and the larger the generalizability.

Perhaps because of their complexity and the lack of support in standard statistical programs, G-analyses are rarely used (for exceptions see Cain & Green, 1983; Doverspike, Carlisi, Barrett, & Alexander, 1983; Hollander & Harvey, 2002). In addition, a G-analysis is not appropriate in all situations because it requires multiple items (e.g., KSAOs), gives only a single coefficient for all items considered jointly, and cannot be computed in the absence of cross-rater variance. A G-analysis, therefore, can shed light on the sources of variance in KSAO ratings, but it cannot be used to answer the fundamental question in practice, viz., agreement on a single KSAO.

Given the limitations of these measures, indices of absolute agreement are more appropriate for use in job specifications, both because they measure the extent to which two ratings are interchangeable and because they can be computed for a single item. At the simplest level, agreement can be defined as the percentage of identical ratings given by multiple sources (e.g., Hazel, Madden, & Christal, 1964). Although such measures are easily interpreted, they lack a reference for determining the degree of agreement—they are descriptive but not necessarily informative. Of the measures that overcome this limitation (e.g., the Content Validity Ratio, *T* index), the r_{wg} class of measures are generally more easily computed, more interpretable, and more versatile (Lindell & Brandt, 1999). The r_{wg} indices can be computed for a single item, make no assumption about cross-item or cross-rater variance, and, most importantly, reflect absolute agreement in terms of both the pattern and level of ratings. This latter property, as noted by Harvey and Wilson (1998), is essential for studies of KSAO rating agreement because the concern, in practice, is not only which KSAOs are important, but also at what level they are needed.

The logic behind r_{wg} coefficients is to compare the observed variance (and hence agreement) to the variance expected if raters responded according to some hypothetical pattern (e.g., randomly). Finn (1970), for a single item, defined r_{wg} as one minus the ratio of observed variance to the population variance of the uniform distribution having the same number of scale points as the scale used to make the ratings. James et al. (1984) extended this index to cover multiple items by using the average item variance in the numerator and applying the Spearman-Brown correction. However, as noted by Lindell et al. (1999), the application of the Spearman-

Brown correction is not appropriate because r_{wg} is a measure of agreement, not reliability. That is, although lengthening a test by adding parallel items will increase reliability, doing so will not increase agreement. Nonetheless, Cohen, Doveh, and Eick (2001) reasoned that lesser agreement on many items might be equivalent to greater agreement on fewer items. However, given items with equal variances, applying the correction could lead to the untenable situation in which raters manifest essentially no agreement on any one item, but across all items show high agreement. Therefore, r_{wg} is best calculated without the correction by simply using the average of the item variances as the numerator.

Unfortunately, there is no consensus on the most appropriate reference variance. The variances that have received the most attention are the population variance of the relevant uniform distribution (James et al., 1984) and the asymptotically maximum variance (Lindell et al., 1999), neither of which is ideal in many situations. The population variance of the uniform distribution, designed to represent random responding, yields values that are scale-dependent, can take on inadmissible values of less than zero when agreement is worse than random (Lindell & Brandt, 1997), and is a realistic representation of random responding only when there are at least as many raters as scale points and ratings are made without bias or the biases cancel out (Brown & Hauenstein, 2003). Using the maximum possible variance also yields results that are scale dependent (Lindell et al., 1999) and, more importantly, may be too lenient (and unrealistic) because few empirical distributions will ever take the form of bipolarity at the extremes. By comparing the observed sample variance to a theoretical population variance, r_{wg} calculated with either of these variances will be sample size dependent (i.e., values will be upwardly biased for all $N < \infty$). James et al. (1984) recognized this indeterminacy, suggesting that variances from a range of distributions (e.g., with varying degrees of skew) could be used. Other permutations are also possible, such as the sample variance of the uniform distribution, the maximum possible sample variance, the variance of the uniform distribution covering the effective rather than theoretical range of scale values (Harvey & Hollander, 2002), or, when ratings within raters across items are uncorrelated, the reference variance divided by the number of scale points (Lindell, 2001). In short, the appropriate denominator may not be the same from one sample or test to another, and should be rationally chosen based on sample and item characteristics.

To overcome these limitations, Brown and Hauenstein (2003) proposed the a_{wg} index, which is defined as the ratio of observed variance to the maximum possible sample variance at the observed mean. Brown and Hauenstein noted that the use of the maximum possible variance or uniform variance in the calculation of r_{wg} is problematic because each assumes that observed ratings are comprised only of true score variance and random error variance. Recognizing that rating bias also is typically present in ratings, Brown and Hauenstein developed a_{wg} as an analog to Cohen's Kappa. Whereas Kappa controls for guessing, a_{wg} estimates agreement under the worse case scenario in which all the systematic variance reflects shared rater bias. The advantages of the a_{wg} index include that it generally has a range of 0 to 1, is independent of the number of raters, and is independent of the number of scale points. A multi-item version of a_{wg} can be computed by averaging the a_{wg} values of individual items.

The a_{wg} index, though, does have some computational complexities. First, a_{wg} is not defined when the observed mean equals one of the scale midpoints (i.e., the maximum possible variance is 0, which requires division by 0). In such cases, when agreement is perfect, a_{wg} should be set equal to 1. Second, because of the way maximum variance is calculated (viz., as a function of the number of ratings at the scale minimum and maximum), a_{wg} is not interpretable at means beyond a given point. As the number of raters increases, so does the number of possible rating combinations and with them the range of interpretable a_{wg} values. Brown and Hauenstein recommend that a_{wg} be used only when the number of raters is at least one less than the number of scale points. When this sample size condition is met, any observed mean outside the interpretable range likely represents high agreement, although it cannot be quantified in the same way as other a_{wg} values.

Unfortunately, there is no straight-forward way to assess the significance of either a_{wg} or r_{wg} values. James et al. (1984), for example, recognizing that any number of reference distributions could be used, recommended that researchers identify the likely range of the "true" r_{wg} value by comparing r_{wg} computed using the smallest expected reference variance with r_{wg} computed using the largest expected reference variance. Lindell and Brandt (1999), on the other hand, recommended the use of a χ^2 statistic to judge the significance of a single value, and Lindell (2001) advocated the use a dependent-samples *t*-test to examine the difference between two items. As noted by Cohen et al. (2001), however, the use of these inferential statistics are not appropriate because the sampling distribution of r_{wg} (and a_{wg}) is not known. Instead, Cohen et al. recommended the use of Monte Carlo simulations and bootstrapping procedures to estimate significance. Based on such procedures, Brown and Hauenstein (2003) recommended a cutoff of .80 as indicating acceptable agreement using a_{wg}. The recommended cutoff for r_{wg} is .70 (Lindell & Brandt, 1999).

As the above makes clear, the choice of an agreement index is not a simple matter. IRCs have been used most frequently, but they do not, strictly speaking, index agreement. ICCs have been used infrequently and assume that the stringent conditions of classical test theory apply. G-theory overcomes this limitation, but can be complex to compute and cannot be used to assess agreement on a single item. The logic of r_{wg}—comparing observed agreement to some standard—is sound, but it is theoretically a very lenient index. Maximum variance (i.e., bimodal) distributions will almost always be an unrealistic (and therefore inappropriately lenient) referent; and r_{wg} values based on the uniform distribution are distributed as a U, with greater leniency at the scale extremes. In contrast, the a_{wg} index has the clear advantage of being independent of the observed mean. The concept of disagreement is unequivocally operationalized as the maximum possible variance at the observed mean. Thus, a_{wg} values will be higher than r_{wg} values based on the uniform distribution for means at the scale midpoint, but lower for means near the scale extremes. The advantage of a_{wg} is that the meaning of the null distribution is constant for all means. However, a_{wg} has yet to be widely applied, which makes comparisons with existing research difficult. To summarize, then, there is little reason to measure agreement with either IRCs or ICCs. The best strategy for assessing agreement in KSAO ratings is to report the results of r_{wg}, a_{wg}, and G-analyses, thereby establishing comparability to prior research while providing the most accurate and useful assessment of agreement.

Prior Research Findings on the Level of Interrater Agreement

Despite the practical and legal importance of a valid job specification, almost no research has been conducted on levels of interrater agreement in KSAO ratings. The dearth of research on KSAO ratings is surprising given the voluminous literature showing that people have great difficulty making accurate and reliable ratings. The social psychological literature (e.g., Pulakos & Wexley, 1983; Tajfel & Turner, 1986) underscores that ratings are just as much due to the rating context as to the rating stimulus itself. Likewise, the literature on the cognitive processes involved in performance appraisal judgments (e.g., DeNisi & Williams, 1988; Feldman, 1981) elucidates the near inability of raters to ever provide completely accurate ratings. In the job analysis and job specification domains, however, treatments of these biases have to date been largely theoretical in nature (Morgeson & Campion, 1997), with the possible exception of research showing that decomposed rating strategies generally provide better data (e.g., accuracy, agreement) than holistic rating strategies (Butler & Harvey, 1988; Cornelius & Lyness, 1980; Harvey, Wilson, & Blunt, 1994; Sanchez & Levine, 1994). Along similar lines, DeNisi and Shaw (1977) showed that self-ratings of ability shared less than 20% of their variance with objective tests. If raters cannot judge their own ability levels, how can they be expected to estimate the requirements of an entire occupation? The limits of human cognitive processing place an upper bound on the accuracy and interrater agreement of ratings.

However, even if individual raters are able to accurately and objectively convey their own beliefs about worker requirements, there is little reason to believe that other raters will share these beliefs. Meaningful within-title differences in job requirements (Green & Stutzman, 1986; Harvey, 1986; Schmitt & Cohen, 1989; Stutzman 1983) may translate into within-title differences in required KSAOs. That is, every position is unique by virtue of the fact that every incumbent occupies a unique role in the organization (Ilgen & Hollenbeck, 1991; Kahn, Wolfe, Quinn, Snoek, & Rosenthal, 1964). If every position is unique, then every position has its own specific worker requirements, which may or may not differ substantially from those of similar positions in the same job title or occupation. In fact, individual positions within a job and individual jobs within an occupation (not to mention ratings themselves) may differ for any number of reasons, including individual (Borman et al., 1992; Conley & Sackett, 1987; Kerber & Campbell, 1987; Mullins & Kimbrough, 1988; Wexley & Silverman, 1978; Wright, Anderson, Tolzman, & Helton, 1990) and organizational (Lindell et al., 1998) performance, such demographic variables as gender (Landy & Vasey, 1991; Schmitt & Cohen, 1989), race (Landy & Vasey, 1991; Schmitt & Cohen, 1989; Veres, Green, & Boyles, 1991), age (Silverman, Wexley, & Johnson, 1984), and education (Ash & Edgell, 1975; Green & Veres, 1990; Landy & Vasey, 1991; Mullins & Kimbrough, 1988; Sanchez & Fraser, 1992; Sanchez & Levine, 1994), and such affective variables as task liking (Love, Bishop, & Scionti, 1991) and job satisfaction (Conte, Dean, Ringenbach, Moran, & Landy, 2003; Jones et al., 2001). Even the ipsative scales (e.g., relative importance) typically used in job specifications encourage cross-position rating differences (Harvey & Wilson, 2000). Given this multitude of factors, it would be surprising if incumbents within a job or occupation shared the same conceptualization of worker requirements. Although different conceptualizations will be similar, they will likely not be similar enough to yield acceptable levels of interrater agreement. As noted by Harvey and Wilson (2000), there is no reason to believe that incumbents in an occupation will all share the same profile of worker requirements.

Be that as it may, interrater agreement on KSAO ratings has been assessed in only a handful of studies. First, Hughes and Prien (1989) calculated all possible IRCs among eight incumbents and supervisors who rated 100 “job skills” (e.g., ability to read, reaction time) on importance, difficulty to acquire, and where acquired. IRCs ranged from –.03 to .50 (median = .33) for importance, .32 to .64 (median = .53) for difficulty to acquire (excluding three suspect raters), and .43 to .73 (median = .52) for where acquired.

Second, Jones et al. (2001, Study 1) had 36 teachers (i.e., incumbents), 47 students, and 31 practitioners rate the trainability of 22 KSAOs (e.g., conscientiousness, knowledge of development theory). ICCs (they did not specify which model they used) were .93 for teachers and students and .89 for practitioners. Jones et al. reported r_{wg} values of .98 in each of their three rater groups, although the reported variances make this value improbable. Using the original formulation of James et al. (1984) and the reported variances for their group of teachers yields

$r_{wg(MV)}$ = .95 and $r_{wg(EU)}$ = .87. Using the more appropriate formulas of Lindell et al. (1999) yields $r^*_{wg(MV)}$ = .79 and $r^*_{wg(EU)}$ = .58 for teachers, $r^*_{wg(MV)}$ = .76 and $r^*_{wg(EU)}$ = .51 for students, and $r^*_{wg(MV)}$ = .82 and $r^*_{wg(EU)}$ = .64 for practitioners. Thus it is not clear how Jones et al. arrived at their value of .98.

Finally, Cornelius and Lyness (1980) had 115 incumbents rate their jobs on 13 various holistic elements, seven of which can be considered KSAOs (e.g., motor coordination, reasoning). Across each of the 10 jobs included in the study and three rating conditions, average IRCs (calculated across all 13 scales) ranged from –.06 to .83, with a median of .38. Geyer et al. (1989) had expert job analysts rate various jobs on similar scales after observing the work and interviewing job incumbents. Using the r_{wg} formulation of James et al. (1984), Geyer et al. reported a wide range of values, from .50 to near 1.00, with many in the .90s.

Many more studies have examined interrater agreement on job analysis ratings. The primary difference between these ratings and those of KSAOs is in terms of the rating object—the job in the case of job analysis, the worker in the case of KSAOs. As is detailed below, job analysis ratings should generally yield higher agreement than ratings of more abstract worker requirements.

The most-studied single instrument has been the Position Analysis Questionnaire (PAQ; McCormick, Jeanneret, & Mecham, 1972). McCormick, Mecham, and Jeanneret (1977) cited average IRCs on the PAQ ranging from .68 to .84 for job analysts, incumbents, and supervisors, though these values are likely spuriously high, a result of trivial agreement on items that do not apply (Harvey & Hayes, 1986; Harvey & Wilson, 1998). Smith and Hakel (1979) found average IRCs ranging from .49 to .63 for different groups rating 25 jobs. Jones, Main, Butler, and Johnson (1982) reported an average IRC of .48 among college students. Across nine jobs rated with the PAQ, Cornelius et al. (1984) reported average IRCs that ranged from .29 to .85, with a median of about .55. Across 24 jobs rated with the PAQ, DeNisi et al. (1987) reported average IRCs of .72 and .85 for naïve and expert raters, respectively. Surrette et al. (1990) reported an average IRC of about .60 for untrained analysts (primarily college students) who had varying amounts of job knowledge.

Similar levels of agreement have been reported in studies using other instruments. Cain and Green (1983) reported average IRCs and generalizability coefficients, calculated on ratings of generalized work behaviors of occupations in the *Dictionary of Occupational Titles*, in the range of roughly.50 to .90. Sanchez and Levine (1989) obtained average IRCs from incumbents in four jobs that ranged from .14 to .26 (median = .20) for task time-spent and from –.02 to .40 (median = .28) for task importance. Pine (1995) reported average IRCs of .38 on an absolute time-spent scale, .33 on a relative time-spent scale, and .50 on an importance scale. Hughes and Prien (1989) reported IRCs in the range of .30 to .40 for ratings of task importance. Harvey and Hollander (2002) computed average IRCs on 1147 "occupational units" in the O*Net abilities questionnaire database. At the level of rater pairs, IRCs ranged from –.46 to .99 (median = .51); at the aggregate level, IRCs ranged from .04 to .87 (median = .48). Across six jobs, Sanchez and Fraser (1992) reported median ICCs (they did not name the model they used) ranging from .49 to .96, with medians of .82 for time spent ratings, .86 for ratings of difficulty to learn, .76 for criticality ratings, and .75 for importance. Manson et al. (2000) reported agreement ICCs ranging from .68 to .89 for the job of Fire Lieutenant and from .78 to .94 for the job of Communications Technician. Using r_{wg}, Lindell et al. (1998) found values ranging from .22 to .57 (median = .45) across eight scales. Harvey and Hollander (2002) computed several different versions of r_{wg} by varying the reference variance, and found, for example, median values of $r_{wg(EU)}$ = .81 and $r_{wg(MV)}$

= .92 using the theoretical eight-point distribution, and $r_{wg(EU)}$ = .50 using an empirically appropriate five-point distribution (i.e., three of the eight scale points were very rarely used). Hollander and Harvey (2003), similarly, reported mean $r_{wg(EU)}$ values of .54 and .76 for O*Net ratings of general work activity importance and level, respectively, generalizability coefficients of about .90, and average IRCs of about .50. Estimates of $r_{wg(EU)}$ can be calculated in two other studies (Borman et al., 1992; Wexley & Silverman, 1978) using the reported standard deviations, and are similar to those reported by Lindell et al. (1998). Dierdorff and Wilson (2003) meta-analytically combined 214 IRCs from 31 studies, finding mean sample-weighted estimates of .77 for tasks and .61 for generalized work behaviors. Coefficients of stability, calculated on 85 coefficients from 15 studies, were .68 for tasks and .73 for generalized work behaviors. These estimates dropped to below .40 when recomputed assuming 5 raters and 100 items. Even these levels may be spuriously high, however, given that the majority of coefficients used in this study almost certainly were originally calculated without removing “does not apply” responses (Harvey & Hayes, 1986; Harvey & Wilson, 1998).

Similar levels of agreement—and similar variability—have been reported in the job evaluation (e.g., Hahn & Dipboye, 1988) and performance appraisal literatures. Conway and Huffcutt (1997), for example, reported meta-analytic mean IRCs of .30 for performance ratings given by subordinates, .50 for ratings given by supervisors, and .37 for ratings given by peers. Although these findings do not imply that agreement must be low in KSAO ratings (e.g., a group of peers would be expected to observe different facets of a supervisor’s performance), they illustrate that disagreement is a nearly universal phenomenon in subjective rating tasks.

Overall, three conclusions seem warranted. First, interrater agreement often fails to reach acceptable levels (e.g., r_{wg} > .70). The more SMEs disagree, the less applicable a mean rating will be to any one position. Selection systems, for example, may as a consequence lead to the hiring of employees whose KSAOs do not match the requirements of the job. Second, the majority of studies have used measures of agreement, such as average IRCs, that inappropriately index agreement. Thus, most conclusions of levels of agreement have been based on faulty measures, making it difficult to know the true extent to which different raters provide interchangeable ratings. Third, estimates of agreement vary widely from study to study and sample to sample. Overall, then, it appears as though SMEs will often disagree in their ratings of KSAO importance, although it is not possible to estimate a precise "normative" level of agreement, and there appears to be as yet unknown factors that moderate the level of agreement. Knowing these factors will help researchers and practitioners design rating materials that maximize interrater agreement, a necessary but not sufficient condition for the validity of the ratings.

Sources of Interrater Disagreement

There is no reason to believe—and no evidence to suggest—that SMEs will agree or disagree at a constant level from one rating task to another. Rather, the findings reviewed above suggest the presence of moderators that influence the extent to which different SMEs will provide interchangeable ratings. As in all rating tasks, a potential moderator can be categorized as either a characteristic of the rater, a characteristic of the rating stimulus, or a characteristic of the rating task (i.e., the interface between rater and stimulus). Figure 1 shows a conceptual model of sources of disagreement. This research focuses on three sources, the rater characteristic of occupational tenure, the stimulus characteristic of occupational complexity, and the rating task characteristic of KSAO abstractness.

Occupational Tenure

Occupational tenure refers to the length of time an incumbent has been employed in an occupation. Beyond common-sense admonitions that incumbents should have some experience (e.g., six months) with the job, little is know about the quality of ratings from less vs. more experienced incumbents. If less experienced incumbents provide low-quality ratings, then practitioners should not include them in rating panels; on the other hand, if tenure does not impact rating quality, then the job specification process would be simplified by justifiably including any available incumbent.

The concept of role making (Graen, 1976; Graen & Scandura, 1987; Kahn et al., 1964) suggests that new employees first sample a variety of tasks and behaviors, before eventually settling into a relatively well-defined role. Given adequate autonomy and task variety, the tasks and duties that comprise an employee's role will change over time, and with it knowledge of the job. Several inexperienced raters may each have different levels of knowledge and may know different aspects of the job. Experienced workers, on the other hand, are more likely to have knowledge of the full range of possible duties and the worker characteristics needed to perform them. This shared job knowledge should result in a common conceptualization and stereotype of the job, which will provide a common frame of reference when making ratings. In other words, experienced incumbents are more likely than inexperienced incumbents to be rating the "same" occupation, which is expected to yield greater interrater agreement.

Most of the research on tenure has focused on the degree to which those with different levels of experience produce similar mean ratings. Goldstein, Noonan, and Schneider (1992, cited in Tross and Maurer, 2000) and Tross and Maurer (2000; 2002) have shown that more experienced raters give higher KSAO importance ratings, and there is some indication that experience can predict job analysis ratings (Arvey, Davis, McGowen, & Dipboye, 1982; Borman et al., 1992; Landy & Vasey, 1991; but see Cornelius & Lyness, 1980; Mullins & Kimbrough, 1988; Schmitt & Cohen, 1989; Sanchez & Fraser, 1992).

Little is known, however, about the effect of experience on interrater agreement. In fact, to my knowledge no one has ever explicitly examined tenure as a moderator of agreement. However, two studies have reported standard deviations of ratings from groups with different levels of experience. First, Tross and Maurer (2000) collected ratings from 209 managers who were categorized into three levels of experience: less than 1 year, 2 – 6 years, and 7 or more years. Experience appeared to have little consistent effect on ratings of KSAO components, though on ratings of task frequency the standard deviations were lowest for the low tenure group—indicating that greater experience was associated with less agreement. In the second study, Borman et al. (1992) obtained task time-spent ratings from 580 stockbrokers. The incumbents were separated into three groups: least experienced (less than 1 year), less experienced (1 – 4 years), and more experienced (more than 4 years). There was no consistent pattern of standard deviations across the 12 tasks rated, indicating no discernable effect of experience. Clearly, more research is needed.

Although somewhat tangential, a comparatively large literature shows that providing more job relevant information generally increases agreement, suggesting that experienced employees, who have more job knowledge, will agree more than inexperience raters. Harvey and Lozada-Larsen (1988) concluded that those with reduced job information are not accurate enough to replace job experts. DeNisi et al. (1987), for example, found higher IRCs on the PAQ for college students rating a job they had studied throughout the semester than those rating from a job title only. Friedman and Harvey (1986) reported a similar trend for students rating from a job title only vs. those rating with descriptive information. Surrette et al. (1990) also found such a trend on PAQ and JCI ratings obtained from college students. Hahn and Dipboye (1991) found that providing more information to otherwise naïve college students resulted in more reliable job evaluation ratings, especially for those who received training. Cornelius et al. (1984; see also Smith and Hakel, 1979) found that IRCs were higher among incumbents and job analysts than among college students, and that among college students agreement was correlated *r* = .58 with a measure of familiarity with the job rated. The standard deviations reported by Jones et al. (2001) show that professional job analysts agreed the most, followed by teachers (incumbents) and then students. Research, therefore, supports the commonsense notion that agreement is positively related to the amount of job information available (Harvey, 1991).

These studies, which show that increasing the information available to raters results in greater agreement, provides evidence against the shared stereotypes hypothesis (Smith & Hakel, 1979), according to which naïve raters, because they share a common stereotype of the occupation being rated, should agree as strongly as knowledgeable raters. Although naïve raters will often posses similar job stereotypes (Burnkrant, 2000; Paunonen & Jackson, 1987; Reed & Jackson, 1975; Rothstein & Jackson, 1980), the similarity does not appear sufficient to produce the level of agreement required for valid KSAO ratings. Once on the job, however, incumbents will be exposed to similar experiences and over time may, as a result, develop similar knowledge structures of the job and worker requirements. Experienced incumbents are more likely to draw on similar expectancies when making their ratings, and should therefore show higher interrater agreement.

Occupational Complexity

Complex occupations are characterized by the absence of set procedures and guidelines for performing work activities, autonomy from direct control, and work outcomes that have a widely applicable and important impact on others. Compared to those in simple occupations, those in complex occupations have greater latitude in determining the specific tasks and duties they must perform to accomplish what are often vaguely defined work objectives. In contrast, the work in simple occupations can be performed by a small set of circumscribed activities. If agreement is less likely in complex than simple occupations, then practitioners may be forced to apply different standards of rating quality to different jobs.

The behavioral ambiguity of complex occupations implies three potential sources of disagreement in ratings: (1) ratings accurately reflect the requirements of an incumbent's unique role; (2) ratings reflect the application of role-specific stereotypes of worker requirements; and (3) cognitive limitations and biases introduce error variance into individual ratings.

The first potential explanation draws on role theory (Graen, 1976; Graen & Scandura, 1987; Kahn et al., 1964) and assumes that raters make accurate, behavior-based ratings. Generally speaking, the work in complex occupations can be expressed in a variety of ways, which manifests as a variety of roles. Where multiple roles exist, so do multiple profiles of required worker characteristics (Harvey & Wilson, 2000). In simple jobs with only a few circumscribed roles, a small number of highly similar profiles may sufficiently describe the entire domain of worker requirements. A greater range of profiles would be needed to cover the requirements of a complex occupation. Thus, when estimating worker requirements, those in simple occupations will be more likely to recall the same work behaviors, whereas those in complex occupations will be more likely to recall different work behaviors. Those in simple occupations will base their ratings on the same behaviors, and so should agree; those in complex occupations will base their ratings on different behaviors, and so should disagree. This explanation, however, requires the almost certainly false assumption (e.g., Richman & Quinones, 1996) that ratings will based on the recall of specific events.

The second potential explanation also builds off of role theory, but assumes that SMEs develop and base their ratings on occupational stereotypes rather than on recall of specific behavioral events. If incumbents perform different activities, then they will develop unique mental representations of the work. Following the logic of the previous paragraph, these different conceptualizations will result in rating disagreement. In actual rating tasks, however, the use of stereotypes and the result of doing so will depend on the complexity of the occupation and the degree of overlap among the stereotypes of different incumbents. The limitations of human cognition (e.g., Schneider, Dumais, & Shiffrin, 1984) suggest that incumbents will be more likely to use stereotypes when the rating task is complex. Given that occupational complexity translates into rating complexity, incumbents in complex occupations will be more likely than those in simple occupations to rate based on their stereotypes. In simple occupations, the use of stereotypes may be largely irrelevant because incumbents will share the same behavioral histories: Ratings from recall will be similar because incumbents will be recalling the same events; Ratings from stereotypes will be similar because the stereotypes are themselves similar. In complex occupations, on the other hand, behavioral histories and stereotypes need not overlap to a great degree. In such a case, if ratings are based on recall (which seems unlikely), SMEs will likely give different ratings. If ratings are based on stereotypes (which seems more likely, especially when rating abstract constructs like KSAOs), the extent of interrater agreement will depend on the similarity of one stereotype to another. As noted above, although people share similar occupational stereotypes (e.g., Paunonen & Jackson, 1987), the similarity may not be enough to yield interchangeable ratings. All else being equal, stereotypes of complex occupations should be less similar than stereotypes of simple occupations. As a result, agreement should be lower in complex than simple occupations.

A third potential explanation—which is not independent of the others—is that information about complex occupations will be more difficult to process than information about simple occupations. Judging a complex stimulus—attendant with more information to be integrated—should exacerbate the problems inherent in making a rating (e.g., of an AO vs. a task) that is itself complex. Cornelius and Lyness (1980), for example, reported some evidence that a decomposed strategy worked better for more complex jobs, but that a holistic strategy worked just as well for simpler jobs. Although more information may theoretically be available with which to make a judgment, it may function as noise that distorts the rating task. In other words, the additional information provided by complex jobs will hurt, not help agreement, independent of the degree to which different incumbents perform similar activities.

Unfortunately, to my knowledge, no one has yet explored the effect of occupational complexity on interrater agreement in KSAO or job analytic ratings, making it difficult to form firm predictions of its effect on ratings. However, Conway and Huffcutt's (1997) meta-analysis on multisource performance ratings provides some evidence that agreement may be lower for complex occupations. They found that supervisor ratings and peer ratings showed higher "interrater reliability" in nonmanagerial as opposed to managerial occupations. More precisely, they reported IRCs of .60, .52, and .48 for supervisor ratings and .31, .43, and .41 for peer ratings, for low (e.g., skilled), medium (technical), and high (e.g., professional) complexity occupations, respectively. Conway and Huffcutt, along with Harris and Schaubroeck (1988), also reported evidence that cross-source agreement was higher for less complex occupations. Because complex jobs involve a range of often disjointed and infrequent activities, raters may have neither the ability nor the opportunity to observe the work in all its facets. Different raters, as a consequence, will likely have, in the case of performance appraisal, different conceptions of the work performed and, in the case of setting worker requirements, different notions of the KSAOs needed for successful performance.

The Interaction of Occupational Tenure and Complexity

According to the arguments advanced above, occupational tenure will tend to increase interrater agreement by setting a common frame of reference and by enabling an incumbent to understand the full range of potential job behaviors. Occupational complexity, on the other hand, will tend to decrease interrater agreement by making it more difficult for raters to integrate all the available information and because different raters may perform the same job in different ways. Given these effects, tenure and complexity are likely not additively related, but rather interact to influence agreement.

Because simple occupations comprise only a small number of roles, new hires can quickly explore and adopt them. In addition, because the work is easily learned, new hires can reach proficiency very quickly. Complex occupations, in contrast, comprise many multifarious roles, and the required duties may take years to completely master. In some instances, new employees may not even be allowed to engage in some functions (e.g., preparing budgets or managing projects). As employees gain experience, though, they will become more familiar with the full range of occupational duties, and thereby develop a fuller conceptualization of occupational requirements in term of the associated KSAOs. Agreement, therefore, should be relatively high in simple occupations, even for new hires, though greater experience will still tend to increase agreement. Agreement among inexperienced employees in complex jobs should be relatively low, but relatively high among those with longer tenure.

KSAO Abstractness

A third factor that may influence interrater agreement in ratings of worker requirements is the abstractness of the KSAOs themselves. As Harvey (e.g., Harvey & Wilson, 2000) has pointed out, job specifications often involve direct ratings of abstract, single-item worker characteristics, a rating process that is unlikely to produce high-quality data given the limitations of human cognition and the potential ambiguity of item interpretations. If so, then direct estimation methods (Morgeson & Campion, 2000), such as those on which the O*Net is based and those typically used in competency modeling, must be called into question.

Unfortunately, this assertion is largely theoretical because most of the supporting evidence is only tangential. Dierdorff and Wilson (2003), for example, reported mean IRCs of .77 for task-level job analysis data and .61 for more abstract generalized work activities. Murphy and Wilson (1997) reported similar job-analytic results. DeNisi and Shaw (1977) found that self-ratings of ability correlated below r = .40 with objective ability tests. Research on holistic vs. decomposed job analysis rating strategies has almost always shown that interrater reliability is higher for decomposed than holistic ratings (Butler & Harvey, 1988; Cornelius & Lyness, 1980; Sanchez & Levine, 1994), and studies have shown that holistic ratings correlate quite poorly with decomposed ratings, with *r*s ranging from .20 and less (Butler & Harvey, 1988) to only near .70 (Harvey et al., 1994). Along the same lines, Wanous, Reichers, and Hudy (1997; see also Nagy, 2002; Wanous & Hudy, 2002) reported a mean observed correlation of .63 between single-item and multi-item job satisfaction scales. Although these authors interpret their results as indicating that single-item measures may be acceptable, measures sharing less than 40% of their variance are likely to lead to different conclusions (e.g., Oshagbemi, 1999). The authors acknowledge, however, that single-items are most appropriate for factual items (e.g., demographics), may be appropriate for simple, unitary constructs (e.g., overall job satisfaction), but probably are not appropriate for assessing complex phenomena. By extension, direct estimation of single-item worker requirements will be more justified for concrete attributes, such as KSs, than for abstract attributes, such as AOs. Whereas KSs *describe* observable and verifiable entities, AOs *represent* latent constructs, the characteristics of which must be inferred by the rater. Thus, the existing evidence suggests that, compared to ratings of AOs, ratings of KSs should yield better accuracy, reliability, and interrater agreement.

However, with regards to interrater agreement, the opposite prediction can also be made: interrater agreement should be lower on concrete, specific items. If a KSAO is too specific and too narrowly defined, then ratings may reflect trivial, albeit real, variations in worker requirements (Curnow, McGonigle, & Sideman, 2003; Jeanneret, Borman, Kubisiak, & Hanson, 1999). If true cross-position differences exist, then these will be reflected in ratings of concrete, specific KSAOs. More general, abstract items, on the other hand, will tend to obscure such differences. All else being equal, agreement should be higher on abstract than concrete items. Agreement should be lower on specific items, especially when, as is the case with complex occupations, true cross-position variance exists.

OVERVIEW AND HYPOTHESES

The present study was conducted using KSAO ratings collected as part of two large-scale occupational analyses of clerical, technical, professional, and administrative occupations in the Federal government. Clerical and technical incumbents rated the importance of 31 general KSAOs, and professional and administrative incumbents rated the importance of 44. The incumbents reported their occupational tenure and grade level. The size of these datasets, representing more than 50,000 incumbents in nearly 150 diverse occupations, provided a unique opportunity to test the following hypotheses on a scale seldom seen in the literature.

First, there is little reason to expect incumbents to agree strongly in their ratings, and little evidence to suggest that they do (e.g., Dierdorff & Wilson, 2003; Harvey & Hollander, 2002). Given the probable existence of position-level differences in KSAO requirements, in addition to the complexity of the rating task per se, agreement on ratings of KSAO importance will fail to reach acceptable levels.

Hypothesis 1: Agreement will be lower than conventionally acceptable levels.

Second, because having more information tends to increase rating accuracy and agreement (e.g., Friedman & Harvey, 1986), any factor that provides additional information to raters should result in better agreement. Experience on the job allows incumbents to sample and observe the wide range of potential job activities, and thus develop a broader conceptualization of overall worker requirements. Similarly, sharing the same experiences will likely foster the development of stereotypes that are more similar from one experienced incumbent to another than from one new hire to another. Thus, whether ratings are based on specific, detailed information or on stereotypes, those with longer occupational tenure should agree more in their ratings than those with less experience.

Hypothesis 2: Occupational tenure will be positively related to agreement.

Third, complex occupations comprise a greater variety of potential activities, activities that are themselves more complex, than those in simple occupations. As a result, incumbents in simple occupations, but not necessarily those in complex occupations, are likely to perform similar activities and share similar behavioral histories. If raters use simple behavioral recall when inferring worker requirements, then those in complex occupations are likely to base their ratings on different behaviors, resulting in cross-rater disagreement. If, on the other hand, SMEs base their ratings on their stereotypes or general impressions of their work, one would still expect greater disagreement for complex than simple occupations because stereotypes are more likely to overlap among incumbents in the latter than in the former. In addition, because humans have limited cognitive resources with which to integrate large amounts of information, complex rating tasks typically produce ratings that are less accurate and more variable than are ratings from simple rating tasks (e.g., Butler & Harvey, 1988). For this reason, too, one would expect stronger agreement in simpler occupations. Thus, each explanation suggests that agreement will be higher for simple than complex occupations.

Hypothesis 3: Occupational complexity will be negatively related to agreement.

Fourth, because occupational tenure influences the amount of information available to an incumbent, and because occupational complexity constrains the amount of information available, occupational tenure and occupational complexity should interact to affect interrater agreement. In simple occupations, new hires more quickly learn the full scope of the work and therefore more quickly develop the same conceptualization of worker requirements. Incumbents in complex occupations will generally require more time before understanding the full range of work or developing a common stereotype of worker requirements. In other words, low tenure will have a stronger effect in complex than simple occupations.

Hypothesis 4: Occupational tenure and complexity will interact such that low tenure will result in poorer agreement for complex occupations than for less complex occupations.

Fifth, some KSAOs are more complex and abstract than others, suggesting that item characteristics can influence interrater agreement. To the extent that abstractness obscures true differences, interrater agreement should be higher on abstract items than on more concrete items (Curnow et al., 2003; Jeanneret et al., 1999).

Hypothesis 5: Agreement will be stronger on abstract than concrete KSAOs.

Sixth, in the absence of cross-position differences in worker requirements, specific KSAOs will pick up no more information that more abstract KSAOs. Only when differences do exist will agreement be lower on concrete KSAOs.

Hypothesis 6: KSAO abstractness and occupational complexity will interact such that agreement will be lower for concrete KSAOs and complex occupations than for concrete KSAOs and simple occupations, and will not differ between complex or simple occupations for abstract KSAOs.

METHOD

Datasets

Clerical and Technical Occupations

The data on clerical and technical (CT) occupations were collected from November 1993 through January 1994 as part of a government-wide occupational study (Rodriguez, Usala, & Shoun, 1996). At the time of the study, over 500,000 employees worked in the 77 (54 clerical, 23 technical) occupations targeted. Of the 71,437 surveys delivered to incumbents, 28,575 (40%) were returned and usable. As reported by Rodriguez et al., the sample is representative of the population. Incumbents were in grades GS-3 to GS-9 and rated their own position only. The KSAO list was embedded within a larger survey that included a task inventory and an organizational climate survey.

Professional and Administrative Occupations

The data on professional and administrative (PA) occupations were collected in two waves in 1996 as part of a government-wide occupational study (Pollack, Simons, Patel, & Gregory, 2000). The first wave targeted occupations at the GS-5 level; the second wave targeted occupations at the GS-9, GS-11, GS-12, and GS-13 levels. Only the second wave data was used in the present study because part of the first wave ratings were made by higher-grade incumbents asked to rate at the GS-5 level. Of the 91,633 surveys mailed, a representative 33,720 (37%) were returned and usable. As in the CT study, incumbents rated their own position only. The KSAO list was embedded within a larger survey that included a task inventory and an organizational climate survey.

Measures

KSAOs

In both the CT (see Shoun, 1995) and PA (see Church & Sher, 1998) studies, the KSAOs were identified by first reviewing the organizational and psychological literatures, existing job analyses, and other documents such as position descriptions from both public and private organizations. The resulting lists of competencies (i.e., KSAOs) were matched to existing competency models, and personnel psychologists vetted the list for redundancy and comprehensive coverage. The KSAOs were defined broadly to be applicable to many CT and PA occupations across the public and private sectors. Through a series of focus groups with incumbents, supervisors, and personnel psychologists, the initial lists were reduced to 31 CT KSAOs (Appendix A) and 44 general PA KSAOs (Appendix B). (Thirty-five technical PA KSAOs were also identified, but were not considered here to maximize the applicability of KSAOs across occupations.) In both studies, respondents rated the *importance* of the KSAOs on a five-point scale (1 = Not important, 2 = Somewhat important, 3 = Important, 4 = Very important, and 5 = Extremely important). The CT study, but not the PA study, included a "competency not needed" option.

Occupational Tenure

In both the CT and PA studies, incumbents reported the years and months they had been employed in their current occupational series. Tenure in months was obtained by converting years to months and then adding the two values together. To group incumbents based on tenure, incumbents were categorized into low, medium, and high tenure groups. Following Borman et al. (1996) and Tross and Maurer (2002), low tenure was defined as 12 or fewer months on the job, medium tenure as 13 to 48 months, and high tenure as 49 or more months.

Occupational Complexity

Occupational complexity was operationalized as *grade level*. Because, in the Federal government, positions are classified into grades independent of occupational series, the meaning of a grade is the same for every job. Grade level is determined with OPM's point-method job evaluation, the Factor Evaluation System (FES), using nine factors: (1) knowledge required of the position, (2) supervisory controls, (3) guidelines, (4) complexity, (5) scope and effect, (6) personal contacts, (7) purpose of contacts, (8) physical demands, and (9) work environment. The higher a position rates on these factors, the higher the resulting grade.

Although grade is a function of more than job complexity alone, the factors related to complexity (primarily factors 1 through 5) carry more weight than the other factors (e.g., factors 2 through 5 have a combined maximum point value of 2,200, whereas factors 8 and 9 have a combined maximum of 100). As a result, grade level is primarily determined by job complexity. Nevertheless, an empirical justification for operationalizing complexity as grade level is in order. Accordingly, I computed complexity scores for 81 of the series-grade combinations available in the CT and PA datasets using the dataset of Schay, Buckley, Chmielewski, Medley-Proctor, and Burnkrant (2001), who examined the factors structure of the FES. The analyses conducted by these authors, based on 1554 position descriptions (each of which was vetted for accuracy and validity by OPM classification experts) of a representative sample of 37 occupations, revealed a four-factor structure, one of which, labeled *Job Controls and Complexity*, comprised factors 2 through 5. Complexity scores were calculated as the average of standardized scores on these four factors. These scores were then applied to the corresponding series-grades in the CT and PA datasets. The correlation between grade and this measure of complexity was $r = .95$. Thus, factors other than complexity accounted for only about 10% of the variance in grade level. Given this high correlation, and that this factor-analytically derived measure of complexity is available for relatively few series-grades in the CT and PA datasets, only grade level will be used as the measure of occupational complexity. Note that complexity has the same meaning in all series—i.e., complexity does not differ from series-to-series because each contains the same grades.

KSAO Abstractness

KSAO abstractness was operationalized in two ways, as the abstractness of the construct being rated and as the multidimensionality of the KSAO definition. Both properties relate to the number of different ways a KSAO can be construed. Whereas the former relates most strongly to stereotype-driven ratings, the latter relates more strongly to purposive ratings. In other words, ratings driven by the concept only will be most affected by its abstractness; ratings driven by the definitions will be most affected by their multidimensionality. Seven doctoral-level personnel psychologists, familiar with the CT and PA competencies, rated each item on the abstractness of the construct it represents and the multidimensionality of its definition using a scale scored as 1 = Not at all, 2 = Slightly, 3 = Moderately, 4 = Very, and 5 = Completely. Higher scores indicate greater abstractness and multidimensionality.

Average interrater correlations were .75 and .68 for CT abstractness and multidimensionality, and .72 and .73 for PA abstractness and multidimensionality, indicating that the raters ranked the items similarly. Because the level of abstractness was important, not just relative standing, r_{wg} and a_{wg} were examined to determine acceptable ratings (see Tables 1 and 2). Across all 150 item ratings, r_{wg} and a_{wg} were acceptable in 65% of the cases. Inspection of the data showed that when r_{wg} and a_{wg} led to different conclusions, on all but three items adjusting a single high or low rating by a single scale point was sufficient to bring both into concordance. Therefore, the criterion for acceptable ratings was having an acceptable r_{wg} or a_{wg}, which occurred on 77% of the items. Agreement was acceptable on both abstractness and multidimensionality for 22 (71%) of the CT items and for 22 (50%) of the PA items. Because the raters were not available for discussion to resolve disagreements, an alternative to using only those items on which the rater agreed is to delete outlying ratings to produce agreement. Doing so yielded a correlation of r = .99 between the original and modified mean profiles, with an average difference in means of .30. Using the raw ratings, correlations between abstractness and multidimensionality were r =.09 for the CT items and r = .40 for the PA items.

Preliminary Analyses

Data Cleaning

Clerical and Technical. The CT dataset was cleaned by first deleting cases that had missing values for more than 25% of the KSAO ratings (n = 912). The pattern of missing values for those above this value tended to show a non-random fatigue effect (i.e., missing values were far more likely on the KSAOs that appeared later in the list); below this value, missing values were far more likely to be disbursed randomly throughout the KSAOs. Removing these cases had virtually no effect on the data: the maximum difference between means was .004 and the maximum difference between standard deviations was .001. Missing values were not imputed because they would not be in practice. Next, to obtain stable estimates of interrater agreement, only those series-grade combination with $n \geq 20$ were retained for analysis. This size is generally considered sufficient to compute bootstrapped estimates of r_{wg} (Cohen et al., 2001). To simplify the analyses and to establish consistency in terms of the complexity levels represented by each occupation, only those CT occupations represented at the GS-4, GS-5, GS-6, GS-7, and GS-8 grade levels were retained for analysis. The final dataset included ratings from 15,039 incumbents in 17 occupations and 85 occupation-grade combinations (see Table 3). “Does not apply” ratings were treated as missing and were not analyzed because of their infrequency (an average of only 8% of the ratings of a KSAO).

Professional and Administrative. The PA data were cleaned following the same procedures used to clean the CT data. Similar to what was seen in the CT data, missing values in many cases appeared non-random (more likely toward the end of the KSAO list) for those with more than 10% missing data. These cases (n = 856) were deleted. The maximum change in mean was .10 and the maximum change in standard deviation was .04, suggesting that removing these cases had minimal impact. Missing values were not imputed. Deleting those series-grades with fewer than 20 ratings resulted in 31,918 ratings. Finally, for the reasons outlined above, only those PA occupations represented at all four available grade levels (GS-9, GS-11, GS-12, GS-13) were retained for analysis. The final dataset contained ratings from 23,004 incumbents in 44 occupations and 176 series-grade combinations (see Table 4).

Assessment of Agency Effects

Because the CT and PA datasets were drawn from occupational analyses, the majority of occupational series contain incumbents from multiple agencies. Grade × Agency multivariate analyses of variances, conducted for each occupation separately and with each KSAO as a dependent variable, showed no effect of agency on mean ratings; i.e., the number of statistically significant agency effects did not exceed chance levels. Although these results do not imply that agreement is the same from one agency to another, they show that, on average, KSAO importance does not vary, which suggests that the meaning of the KSAOs does not depend on agency. Agency was therefore ignored as a substantive variable, because the data were not designed for use at the agency level, because agency had minimal impact on mean ratings, and because doing so simplifies the analyses.

RESULTS

All analyses were conducted separately on the CT and PA occupations. Unless otherwise noted, the CT results are presented first. Given the large number of required statistical tests, the nominal alpha level of .05 was, where multiple tests were performed, adjusted using Holm's (1979) procedure. To conserve space in the tables that follow, series are referenced by series number and KSAOs by item number. The series titles are given in Table 3 and 4 and the KSAOs and their definitions are given in Appendixes A and B. Several supplemental analyses, which did not affect the results, are described in Appendix C.

Hypothesis 1

Hypothesis 1 predicts that agreement will be lower than conventionally acceptable levels. To test this hypothesis, agreement was measured in three main ways. First, r_{wg}, using the population variance of the uniform distribution, was computed for consistency with the currently preferred approach. This index is well behaved under the conditions in the present study (viz., five-point rating scales and large sample sizes; Lindell et al., 1999). In addition, use of the uniform distribution is consistent with the traditional notion of rating variability being due to bias that cancels out across raters. Second, a_{wg} was used because it is scale- and sample-independent and because it is not confounded with mean ratings. Third, G-theory was used to estimate agreement after controlling for irrelevant sources of variance. Values of $r_{wg} \geq .70$, $a_{wg} \geq .80$, and $\Phi \geq .80$ served as the criteria of acceptable agreement.

$\mathrm{r_{wg}}$ *and* $\mathrm{a_{wg}}$

Table 5 summarizes item-level r_{wg} agreement for the CT occupations. For no series-grade does the median r_{wg}, calculated across the 31 CT items, exceed the .70 threshold that indicates acceptable agreement. The maximum r_{wg} is greater than .70 for only 62% of the series-grades, and in many cases the minimum value indicates agreement at or, in one case, below random levels. Looking at all 2,635 item ratings, in only 180 cases (7%) is $r_{wg} \geq .70$. Table 6 shows analogous results for a_{wg}. In no case does median agreement exceed the .80 threshold, and for only 20% of the series-grades is the maximum agreement across the 31 items greater than .80. For only 27 (1%) of the 2,635 ratings is $a_{wg} \geq .80$. The results for the PA occupations are similar. As seen in Table 7, for no series-grade does the median r_{wg} exceed .70. The maximum value is greater than .70 for 93% of the series-grades, but the minimum values are less than 0 for 64% of the series-grades. This greater variance—relative to the CT occupations—may be due to the greater number of PA items (44 vs. 31). Of the 7,744 item ratings, only 950 (12%) have agreement greater than .70. The a_{wg} results, shown in Table 8, indicate comparatively less agreement than found with r_{wg}. For no series-grade does median agreement exceed .80, and for only 13% is the maximum greater than .80. Across all item ratings, agreement is greater than .80 only 54 times (.7%). These results support Hypothesis 1.

Note that these a_{wg} results are based on treating as missing values that are uninterpretable due to an extreme mean and small sample size. Although there were no such cases in the CT dataset, there was 20 (.2%) in the PA dataset. Because many of these uninterpretable values likely represent high agreement (Brown & Hauenstein, 2003), the results shown in Table 8 may underestimate true a_{wg} agreement. To help correct for this, a_{wg} was also calculated after substituting the minimum *n* needed to produce an interpretable result at the given mean. Doing so, however, did not change any of the results.

Tables 5 through 8 also show the percentage of values (i.e., of the 31 CT and 44 PA items) whose bootstrapped 95% CI upper bound is greater than or equal to the threshold for acceptable agreement. These analyses estimate the "best case scenario," and thus are more lenient than the point estimates described above. The analyses are based on resampling 1,000 times and items that received valid ratings from at least 20 incumbents. If agreement could not be computed for a sample (e.g., missing values, an uninterpretable a_{wg}), another sample was drawn and the process repeated until 1,000 valid estimates were obtained. As seen in the tables, the percentage of acceptable agreement varies considerably from series-grade to series-grade. For the CT occupations, r_{wg} agreement is acceptable for as few as 0 and as many as 74% of the items (*Med* = 19%, *M* = 26.09%), and a_{wg} agreement is acceptable for as few as 0 and as many as 90% of the items (*Med* = 6%, *M* = 16.94%). For the PA occupations, similarly, the percentage of acceptable r_{wg} agreement ranges from 5% to 89% (*Med* = 27%, *M* = 31.22%), and the percentage of acceptable a_{wg} agreement ranges from 0% to 86% (*Med* = 7%, *M* = 14.22%). Note that the difference between the medians and the means indicates a strong positive skew—although agreement is high for some series-grades, it is low for most. These results also support Hypothesis 1.

Tables 9 and 10 summarize multi-item agreement for the CT and PA occupations, respectively. Because the multi-item indices are simply the average of the single-item values, the standard deviations are also shown. Given the large sample size, mean agreement is similar to median agreement. As would be expected, therefore, for no CT or PA series-grade did agreement reach acceptable levels. Note that the standard deviations for r_{wg} are larger than those for a_{wg}, a likely result of increased variance due to r_{wg}'s statistical relationship with the mean rating. These results provide further support for the hypothesis.

Although the data were designed for use at the level of grades within series, the low agreement reported above raises the possibility that agreement will be higher for more narrowly defined groups of raters. Accordingly, for each dataset I calculated agreement at multiple levels of aggregation. Four levels were possible with the CT dataset: (1) location within agency within grade within series; (2) agency within grade with series; (3) grade within series; and (4) series. Because work location was not assessed in the PA study, only the latter three groups were possible with these occupations. To maximize consistency, only groups with $n > 4$ at the lowest level of aggregation were included in the analyses; i.e., each level of analysis includes data from a common set of participants. As seen in Table 11, mean agreement increases slightly, and the standard deviation increases substantially, as the level of aggregation decreases. These findings are illustrated in Figures 2 – 5, which show that negative skew increases and the tails grow heavier, with a decrease in aggregation. Thus, higher levels of aggregation are associated with somewhat less agreement, although lower levels of aggregation are associated with greater variability, especially at the extremes of agreement. Importantly, not even at the most narrowly defined level is mean agreement acceptable. These findings provide further support for Hypothesis 1.

Despite the low levels of agreement reported so far, the possibility remains that these may be, nonetheless, inflated due to the use of an inappropriately large reference variance (e.g., because respondents did not use the entire 5-point scale, a uniform distribution does not accurately represent random responding). Alternative measures of r_{wg} agreement can be constructed using estimated population variances to represent random responding. That is, observed agreement is compared to random responding, which is defined as the variance of raters randomly selected from the population of raters. Tables 12 and 13, for CT and PA occupations respectively, show agreement on each item, across series-grades, calculated using four different estimates. The first estimate is based on the overall sample variance of each item. Monte Carlo simulations were also run, drawing 5,000 groups of 20 randomly selected raters. The variance was calculated for each group, and then the median, 25th percentile, and 75th percentile of the 5,000 variances taken as estimates of the population variance. Note that the 25th percentile represents a stringent standard (small variance), whereas the 75th percentile represents a lenient standard (large variance). r_{wg} was then calculated for each series-grade. As seen in the tables, even in the most lenient case average agreement falls far short of acceptable levels. These results, too, support the hypothesis.

Generalizability Analyses

The question addressed by the generalizability analyses is the extent to which item ratings generalize across multiple raters. That is, how many incumbents must provide ratings before one can be confident in their mean item rating? The stronger the agreement among incumbents, the more their ratings will generalize to the universe of potential raters. By identifying other sources of variance in item ratings, one may then isolate the rater effect. In these analyses, item level is just as important as the relative standing of KSAOs. Accordingly, absolute generalizability coefficients, Φ, were estimated.

For each analysis, the variance components were estimated using the minimum norm quadratic method of SPSS's VARCOMP procedure. Restricted maximum likelihood estimation was also attempted, but for some series the large number of cells prevented its use. However, when it could be used, the estimated variance components were nearly identical to those obtained using the minimum norm quadratic method.

For the first analysis, Φ was estimated using a one facet, item × rater random effects model. Raters were treated as a random effect because they were, in fact, sampled randomly. Items were treated as a random effect because an infinite number of KSAOs, variously defined, could have been assessed. The generalizability coefficients were calculated using the formula:

$$\Phi = \frac{\sigma_i^2}{\sigma_i^2 + \frac{\sigma_r^2}{n_r} + \frac{\sigma_{ir,e}^2}{n_r}},$$

where i = item, r = rater, and e = error. The analyses were conducted on those cases with valid ratings for every item. The results for the CT occupations are shown in Table 14. On average, items accounted for 18% of the variance, raters accounted for 30%, and error and the interaction accounted for 52%. Φ was calculated assuming either 5 or 20 raters. Assuming 5 raters, no Φ reaches the acceptable .80 level; using 20 raters, Φ > .80 for 71% of the series. The table also shows the number of raters necessary to achieve a Φ of .80, which was obtained by solving the above formula for n_r and then rounding up. On average, 21 raters would be needed to reach acceptable reliability. The results for the PA occupations, shown in Table 15, indicate a higher level of reliability. On average, item accounted for 45% of the variance, raters for 16%, and error and the interaction for 40%. Thus, the variance for raters is nearly half that found for the CT occupations. Not surprisingly, then, a higher proportion of Φs are greater than .80 when $n_r = 5$, all Φs are greater than .80 when $n_r = 20$, and the average number of raters needed to achieve Φ = .80 is only 6. The CT results, therefore, support Hypothesis 1, although the PA results do not: 6 raters is not an unusual number to include in a job specification study.

For the second analysis, Φ was estimated using a two facet, item × (rater:grade) random effects model. Grade was treated as a random factor because the purpose is to generalize to all levels of complexity, however defined. Because each rater is associated with only one grade but all grades with all items, the rater factor is nested within the grade factor. Thus, variance components were estimated for item, grade, item× grade, and rater:grade. The generalizability coefficients were calculated using the formula:

$$\Phi = \frac{\sigma_{\mathrm{i}}^2}{\sigma_{\mathrm{i}}^2 + \frac{\sigma_{\mathrm{g}}^2}{\mathrm{n_g}} + \frac{\sigma_{\mathrm{ig}}^2}{\mathrm{n_g}} + \frac{\sigma_{\mathrm{r:g}}^2}{\mathrm{n_r}} + \frac{\sigma_{\mathrm{ir:g,e}}^2}{\mathrm{n_r n_g}}},$$

where g = grade and i, r, and e are as before. The analyses were conducted on those cases with valid ratings for every item. Φ was calculated assuming a single grade, i.e., $n_g = 1$. As seen in Tables 16 and 17, for CT and PA occupations, respectively, grade and the item × grade interaction accounted for almost no variance. As a result, the Φs are very similar to those reported above. The estimated number of raters needed to reach Φ = .80 are also similar. Note, however, the anomalous estimates for series 204 (n_r = -104), 305 (n_r = 536), 331 (n_r = 818), and 2005 (n_r = 156), a result of large r:g and error components relative to the item component. Assuming $n_g = 2$ yields the more reasonable estimates of 34, 63, 77, and 57, respectively. This analysis, too, provides mixed support for Hypothesis 1.

For the third analysis, Φ was estimated using a three facet item × grade × tenure × rater(grade × tenure) random effects model. Tenure was treated as a random effect because it represents a continuous variable. The generalizability coefficients were calculated using the formula:

$$\Phi = \frac{\sigma_{\mathrm{i}}^2}{\sigma_{\mathrm{i}}^2 + \frac{\sigma_{\mathrm{g}}^2}{\mathrm{n_g}} + \frac{\sigma_{\mathrm{t}}^2}{\mathrm{n_t}} + \frac{\sigma_{\mathrm{ig}}^2}{\mathrm{n_g}} + \frac{\sigma_{\mathrm{it}}^2}{\mathrm{n_t}} + \frac{\sigma_{\mathrm{r(gt)}}^2}{\mathrm{n_r}} + \frac{\sigma_{\mathrm{irgt,e}}^2}{\mathrm{n_r n_g n_t}}},$$

where t = tenure and all other terms are as before. Φ was calculated assuming a single grade and a single tenure. This model was not computed on the CT occupations because no series had at least five usable cases in each grade × tenure cell. Although missing values could have been imputed to produce a minimum of five usable cases per cell, the high prevalence of "does not apply" ratings rendered this option inappropriate (i.e., the majority of invalid ratings were "does not apply"). For the PA occupations, nine series had at least five valid cases in each cell. The results for these nine series are shown in Table 18. Grade and tenure and their interactions with items contributed almost no variance to the ratings. As a result, the Φs and estimated required n_rs are almost identical to the estimates obtained in the other models. It is reasonable to assume that the results for the CT occupations would have been similar. Thus, this analysis also provides mixed support for Hypothesis 1.

Hypothesis 2

Hypothesis 2 predicts that occupational tenure will be positively related to agreement. Only cells (i.e., series-grade by tenure group) with at least 5 raters were analyzed. For the CT occupations, n = 61 for low tenure, n = 82 for medium tenure, and n = 85 for high tenure. For the PA occupations, n = 91, n = 163, and n = 176, respectively. Tables 19 and 20 show average agreement, across series-grades, for each item by tenure group. Although no strong pattern is evident, for a_{wg} there is a slight trend for agreement to decrease with increasing tenure, a trend opposite that predicted by Hypothesis 2.

The data were first analyzed with multiple regression analyses, in which centered mean importance and its square were entered on the first step, dummy coded vectors representing tenure were entered on the second step, and the importance × tenure interactions were entered on the third step. For the CT occupations, the interactions added significant incremental variance for items 6, 27, and 30, for both r_{wg} and a_{wg}. For the PA occupations, the interaction was significant for items 19 and 40 when using r_{wg}, and for items 1, 2 and 40 when using a_{wg}. However, because the results were the same regardless of whether the interactions were included, they were excluded from all subsequent analyses.

Tables 21 and 22 show the results of regression analyses in which centered mean importance and its square were entered on the first step and dummy coded vectors representing tenure were entered on the second. Overall, the results do not support Hypothesis 2. For the CT occupations, tenure explains significant variance in r_{wg} agreement on items 18 and 31, and the trend is in the predicted direction. However, tenure is not significant for these items using a_{wg} agreement. Using a_{wg}, tenure is significant for items 1, 3, 4, and 15, though the effects are small and not in the predicted direction. Similarly, for the PA occupations, tenure explains significant variance in r_{wg} agreement on item 21, 37, and 40, and the trend is in the predicted direction. These effects, however, are not significant when using a_{wg}. Using a_{wg}, tenure is significant for items 4, 5, 9, 10, 12, 17, 25, and 38, but the trends are not in the predicted direction. Even when tenure is a significant predictor, it has only a small effect (maximum $R^2 = .05$).

For the CT and PA a_{wg} analyses, 128 (2%) and 863 (5%) coefficients, respectively, could not be computed because of an extreme mean and a small sample size. Accordingly, these coefficients were re-computed assuming the smallest n required to yield an interpretable value. The regression analyses were run again, but the results were nearly identical to those already reported, and the conclusions are the same.

Hypothesis 3

Hypothesis 3 predicts that occupational complexity will be negatively related to agreement. Tables 23 and 24, which show average agreement on each item for each complexity level (i.e., grade), tell a somewhat different story for r_{wg} and a_{wg}. When measured with r_{wg}, agreement increases with an increase in complexity. In contrast, when measured with a_{wg}, agreement appears unrelated to complexity. As seen in Tables 25 and 26, the difference appears due to mean importance becoming more extreme with increasing complexity: in general, importance increases for “intellectual” items but decreases for “physical” items. These more extreme scores are reflected in higher r_{wg} agreement.

The data were first analyzed with multiple regression analyses, in which centered mean importance and its square were entered on the first step, dummy coded vectors representing complexity were entered on the second step, and the importance × complexity interactions were entered on the third step. For the CT occupations, the interactions added significant incremental variance for items 8 and 13 when using r_{wg}, and for 12 and 13 when using a_{wg}. For the PA occupations, the interactions were significant for items 8, 10, and 23 when using r_{wg}, and for 8, 10, 23, and 28 when using a_{wg}. However, because the results were the same regardless of whether the interactions were included, they were excluded from all subsequent analyses.

Tables 27 and 28 show the results of regression analyses in which centered mean importance and its square were entered on the first step and dummy coded vectors representing complexity were entered on the second. When controlling for mean importance, occupational complexity has little effect on agreement. For the CT occupations, complexity added significant variance only to item 10, for both r_{wg} and a_{wg}. For the PA occupations, complexity added significant variance to r_{wg} and a_{wg} agreement on items 20, 21, 23, 26, 27, and 29. Complexity was also significant for items 28 and 40, but only for a_{wg}. In each case, the effects, though small, are in the direction opposite that predicted. Thus, the results do not support Hypothesis 3.

For the PA occupations, 20 (.3%) a_{wg} values were out of range (there were none for the

CT occupations). For these items, the coefficients were re-computed assuming the smallest *n* required to yield an interpretable value. The regression analyses were run again, but the results were nearly identical to those already reported, and led to the same conclusions.

Hypothesis 4

Hypothesis 4 predicts that occupational tenure and complexity will interact such that low tenure will result in poorer agreement for complex occupations than for less complex occupations. As before, the analyses were based only on cells with at least five raters. *N*s ranged from 5 to 17 in the CT occupations ($M = 15$) and 12 to 44 in the PA occupations ($M = 36$). Tables 29 and 30 show the incremental R^2s for hierarchical regression models regressing agreement on centered mean importance and its square on the first step, effect coded main effect vectors representing complexity and tenure on the second step, and the complexity × tenure interactions on the third step. For the CT occupations, when using r_{wg}, the interactions add significant variance on items 11, 14, 28, and 31. When using a_{wg}, the interactions are significant for items 11, 14, and 21. For the PA occupations, the interactions are significant for items 32 and 34 when using r_{wg}, and for items 1, 30, 31, 32, and 35 when using a_{wg}.

Tables 31 and 32 show, for these items, the cell means and results of simple effects tests conducted using SPSS's MANOVA procedure (Pedhazur, 1997). For the CT occupations, agreement remains relatively constant for medium and long tenure, across all levels of complexity. However, agreement for low tenure spikes at the most complex level. In other words, agreement is higher for those with low tenure in complex positions than for any other group. The form of the interaction is somewhat different in the PA occupations. As in the CT occupations, agreement is relatively constant across all grade levels for those with medium or high tenure. However, for those with low tenure, agreement at low complexity is as strong or stronger than agreement among those with medium or long tenure; at high complexity levels, however, agreement drops. In other words, in the PA occupations, agreement among those with short tenure is high at low complexity levels but low at high complexity levels. These results provide mixed support for Hypotheses 4, in that the predicted pattern was found for those with short tenure in PA occupations (on a small number of items), but the opposite pattern was found for those in CT occupations (on a small number of items).

For the CT and PA occupations, 128 (2%) and 863 (5%), respectively, a_{wg} coefficients could not be interpreted. The analyses were re-run assuming the smallest *n* required to yield an interpretable values, but the results and conclusions were the same.

Hypothesis 5

Hypothesis 5 predicts that agreement will be stronger on abstract than concrete KSAOs. For the CT occupations, r_{wg} and a_{wg}, respectively, correlated $r = .23$ and $r = .28$ with abstractness and $r = .03$ and $r = .08$ with multidimensionality. For the PA occupations, r_{wg} and a_{wg}, respectively, correlated $r = .02$ and $r = .30$ with abstractness and $r = .12$ and $r = .29$ with multidimensionality. To control for mean importance and to estimate the unique effects of each predictor, the hypothesis was further examined, across all series-grades, by regressing agreement on centered mean importance and its square on the first step, centered abstractness and centered multidimensionality on the second, and the abstractness × multidimensionality interaction on the third step.

Table 33 shows the results using only those items with acceptable agreement of abstractness. For the CT occupations, abstractness has a very small positive effect ($b = .03$ and b = *.01*) and multidimensionality a very small negative effect ($b = -.01$ and $b = -.00$) on r_{wg} and a_{wg} agreement, respectively. Likewise, for the PA occupations, abstractness ($b = .02$ and $b = .00$) and multidimensionality (*b* = -.00 and *b* = .00) have very small effects on r_{wg} and a_{wg} agreement, respectively. Using r_{wg} and a_{wg}, the interactions are significant for both the CT (both *b*s = .01) and PA (*b* = .05 and *b* = .02, respectively) occupations. Plotting the interactions (see Figure 6) reveals a somewhat different form for each, but, in general, shows that agreement is strongest when both construct abstractness and item multidimensionality are high, agreement is weakest when construct abstractness is low but item multidimensionality high, and agreement at moderate abstractness is not affected by multidimensionality. In other words, rating a multidimensional item that represents an abstract construct leads to the strongest agreement and rating a multidimensional item that represents a concrete construct leads to the weakest agreement. These analyses were repeated using all items, both with abstractness scores calculated after removing outliers and with the raw ratings. In each case, the results were largely the same, except that the interactions were not significant for the CT occupations. Given the small effects, this difference is not meaningful. Overall, then, the hypothesis was supported—though with very weak effects—when abstractness was operationalized as construct abstractness, but not when operationalized as item multidimensionality.

Hypothesis 6

Hypothesis 6 predicts that KSAO abstractness and occupational complexity will interact such that agreement will be lower for concrete KSAOs and complex occupations than for concrete KSAOs and simple occupations, and will not differ between complex or simple occupations for abstract KSAOs. The hypothesis was tested separately for construct abstractness and definition multidimensionality, across all series-grades, by regressing agreement on centered mean importance and its square on the first step, centered abstractness and effect coded vectors for grade on the second step, and the abstractness × grade interactions on the third. The analyses were repeated using only those items with acceptable agreement on abstractness, all items after outlying abstractness ratings had been removed, and the raw ratings. However, for no model did the interactions explain significant variance. Therefore, Hypothesis 6 was not supported.

DISCUSSION

Review of Results

Level of Agreement

The clearest finding of this research, consistent with Hypothesis 1, is that agreement fails to reach acceptable levels. For no series-grade, the level at which the data were designed for use, did median r_{wg} or a_{wg} agreement reach acceptable levels. At the item level, the percentage of acceptable values ranged from a high of 12% for PA r_{wg} to a low of .7% for PA a_{wg}. In addition, the vast majority of 95% CIs upper bounds of item agreement failed to exceed acceptable levels. Agreement was even worse when r_{wg} was calculated using various reference variances estimated from the population of raters. The low levels of agreement are even more striking when one takes into account that the KSAO lists were constructed to apply to multiple occupations, in multiple agencies, and for multiple grades: some level of agreement should be "built in" because the instrument was not designed to capture molecular differences in worker requirements.

As would be expected, agreement was stronger at lower (e.g., series-grade-agency-location) than higher (e.g., series) levels of aggregation. However, even at the lowest level of aggregation mean agreement failed to reach acceptable levels. Interestingly, both extremely low and extremely high levels of agreement were common at the lower levels of aggregation. Most likely, the larger sample sizes at the higher levels of aggregation obscured these differences. That is, the more stable estimates at high aggregation masked both substantial agreement and disagreement at low aggregation, a phenomenon similar to the aggregation bias (James, 1982).

Thus, increasing the homogeneity of incumbents with respect to organization and geographic location does not necessarily improve agreement.

The G-analyses indicate that a substantial proportion of the variance in ratings is due to rater effects. For the CT occupations, considering raters as the only facet, item accounted for 18% of the variance, the rater main effect for 30%, and the error term for 52%. For the PA occupations, items, raters, and error accounted for 45%, 16%, and 30% of the variance, respectively. The facets of grade level and tenure accounted for essentially 0% of the variance in ratings. Thus, much of the variance in ratings—that attributable to error and higher-order interactions—remains unexplained.

The Φ coefficients indicate substandard agreement for the CT occupations, but adequate agreement for PA occupations. The reason for this discrepancy, especially in light of the fact that agreement measured with r_{wg} and a_{wg} is lower for the PA than CT occupations, is not entirely clear. However, it is conceivable that the greater item variance in the PA occupations obscured the rater effects (the average variance in mean item agreement was .50 for CT series-grades but .96 for PA series-grades). In this case, even if rater variance were constant from item-to-item and the same for CT and PA occupations, the greater variance in item scores for PA occupations would yield a higher Φ. The G-analyses, therefore, present an overly optimist appraisal of interrater agreement in the PA occupations.

Taken together, the results of the r_{wg}, a_{wg}, and G-theory analyses (in addition to the supplemental analysis of average IRCs) are consistent with prior research. That is, average agreement fails to reach acceptable levels, although there is substantial variance in agreement from one occupational unit to another. Importantly, these conclusions stem from the confluence of results based on a variety of agreement indices—indices that are appropriate to the job specification context. Thus, this study provides the most comprehensive examination of interrater agreement to date.

Sources of Disagreement

Little evidence was found for the hypothesis that agreement would be higher among experienced than inexperienced incumbents. Using r_{wg}, tenure had a significant effect in the predicted direction for two CT items and three PA items. However, using a_{wg}, tenure had a significant effect in the direction opposite that predicted for four CT items and eight PA items. All of the effects were small by conventional standards. The weight of the evidence, therefore, points to a small and inconsistent trend for agreement to be lower among experienced incumbents.

In the introduction I argued that experience provides detailed job knowledge that could lead to shared stereotypes of a job. Inexperienced workers, not having such detailed knowledge, could, on the other hand, have very different conceptions of the job. This reasoning relies on the assumption that new hires will perform different work whereas experienced incumbents, if they do not perform similar work, will at least have been exposed to the same types of work. However, the opposite prediction could also be made, viz., that new hires perform similar work and will, as they gain experience, begin to specialize. If such is the case, then agreement would be higher among inexperienced than experienced incumbents. Of course, both forces could be operating simultaneously: some incumbents perform similar work, and some do not, at all levels of experience. In effect, the two would cancel each other out.

Another explanation for the mainly null findings is that tenure only has an effect at very low levels. For example, agreement could begin to level off after one or two months on the job. Similar to Rothstein's (1990) findings on the interrater reliability of job performance ratings, agreement may reach an asymptotic level after a very short time. Unfortunately, the small number of new hires in the present datasets precluded a strong test of this hypothesis.

Hypothesis 3, which predicted that agreement would be lower in more complex occupations, was also not supported. After controlling for KSAO mean importance, complexity had a significant effect on only one CT and eight PA items. In each case, the effect was small and indicated, contrary to the hypothesis, that agreement increases with occupational complexity.

With one exception (Spatial Orientation), each of the eight PA items on which complexity had an effect were managerial KSAOs (e.g., Financial Management, Vision). (Note, however, than complexity did not have an effect on other managerial KSAOs, such as Leadership and Organizational Awareness.) Although in the introduction I argued that complexity would result in worse agreement because different incumbents would have greater latitude to perform different duties, these results suggest that some duties may become more common at higher complexity levels. For example, the technical work of two project managers may be very different, but each would need knowledge of financial management. Borman et al. (1992) reported similar level effects on the job analysis ratings of experienced and inexperienced stockbrokers. Incumbents in a low complexity position may or may not need a particular KSAO that all incumbents in a high complexity position need. Likewise, some KSAOs may be required in low complexity positions, but may or may not be important in high complexity positions. Thus, at all levels incumbents within a single classification will perform different duties, although they may share others.

That said, a molar classification system may have contributed to the null findings. If dissimilar positions and groups of positions are grouped into a single classification, then the inflated variance across positions will obscure variance due to complexity. The classification system used in the present study by necessity fails to make fine distinctions among positions—as a governmentwide system, it must produce a manageable number of series. An examination of more tightly defined jobs could produce clearer results.

Another explanation for the many null findings is restriction in the range of complexity. For example, the difference in FES complexity points between grades 9 and 13 is 1055, compared to a difference of 755 between grades 4 and 8. There may not have been enough variance in complexity to show an effect in the CT occupations.

The fourth hypothesis, that low tenure will result in poorer agreement for complex occupations than for less complex occupations, was partially supported. For the CT occupations, the interaction of tenure and complexity explained significant r_{wg} agreement in four items, and significant a_{wg} agreement in three. For the PA occupations, the interaction was significant for two items when using r_{wg} and five when using a_{wg}. However, the form of the interaction was different for the CT and PA occupations. For the CT occupations, agreement was similar for all grades and all levels of tenure except for grade 8 and low tenure, for which agreement was higher. For the PA occupations, agreement was similar for all grades and medium and long tenure, but for short tenure agreement was higher at low grades than at high grades. In other words, among those with short tenure, agreement increased with complexity in the CT occupations but decreased with complexity in the PA occupations.

The likely explanation for this pattern of results centers around the type of work performed in and the career ladders for CT and PA occupations. In CT occupations, incumbents in grades 8 and 9 are often first line supervisors. In PA occupations, in contrast, grade 9 is entry level and those in grades 13 are, depending on the series, either first line supervisors or senior non-supervisors with much responsibility. The higher agreement in the CT occupations, therefore, may be due to agreement on shared supervisory responsibilities. In the PA occupations, however, the lower agreement at high complexity levels may be among non-supervisors. Unfortunately, because supervisory status could not be assessed in this study, this hypothesis cannot be tested. That said, the interaction was not significant on the majority of items, making any speculation tenuous.

The fifth hypothesis predicted that agreement would be higher on abstract than concrete KSAOs. When abstractness was defined as construct abstractness, a very small effect was found in this direction. However, no effect was found when abstractness was operationalized as definition multidimensionality. The small effect sizes make tenuous any speculation on the origin of this difference, although one may reason that the more concrete items did, as predicted, pick up true cross-position variance, and the incumbents based their ratings on KSAO name more than the context of the definitions. Alternatively, it may simply be the case that the KSAOs represented by the more abstract constructs are truly more common within a job title than are the less abstract constructs, a possibility which cannot be tested without true-scores.

Nonetheless, the interaction between construct abstractness and definition multidimensionality, which was disordinal in the PA occupations, makes interpretation of the main effects difficult. Although the form of the interaction differed by occupational family and the agreement index used, a generally consistent finding was that agreement is highest on abstract multidimensional items, and lowest on concrete multidimensional items. Thus, agreement is lower when incumbents can easily interpret and selectively attend to different aspects of a KSAO definition. When an item is more complex, when abstract and multidimensional, raters may find it cognitively difficult to integrate item content with their behavioral histories and instead resort to rating based on their stereotypes. Be that as it may, the effects were quite small, on the order of only five one-hundredths of a scale point, and so are likely to be unimportant in practical applications. One explanation for these small effects is that item *content* exerts a predominant influence on ratings. That is, construct abstractness and definition multidimensionality may contribute little variance compared to the knowledge, skill, ability or other characteristic described by the item.

The sixth hypothesis—that agreement would be lowest on concrete KSAOs in complex occupations—was not supported. Theoretically, one would predict that easily interpretable items that offer multiple interpretations would capture true cross-position differences, especially in broadly defined occupations that offer a great degree of behavioral latitude. However, in light of the mainly null findings for complexity (Hypothesis 3) and the small effects for abstractness (Hypothesis 5), the lack of an interaction should not be surprising. Here, too, a broad classification or restricted range in complexity, in addition to the predominant influence of item content, may have obscured any effects on agreement.

Limitations

The above conclusions must be tempered with the limitations of the study. First, the available methodology did not allow for a truly comprehensive partitioning of the variance in item ratings. The results reveal substantial variance among incumbents, but grade, tenure, and aggregation level explain little of it. Although G-theory can be used to illuminate such sources of variance, the current design, unfortunately, precluded a truly comprehensive examination. Because raters were nested within the other facets (e.g., grade and tenure), separate effects could not be computed for raters or the two-way interactions. These limitations, though, are not unique to this study: outside of the laboratory, incumbents will always be nested within occupational level, tenure, agency, location, and a host of factors that could potentially account for rating variance. G-theory, therefore, is not as useful in research on job analyses or specifications as it is when fully crossed designs are possible (Shavelson & Webb, 1991).

Second, restricted range in occupational complexity may have masked potential effects. If the items on the CT and PA surveys had been identical, the CT and PA datasets could have been merged to yield complexity scores ranging from grade 4 to grade 13. Such a range would have offered a stronger test of the effect of complexity. A related limitation is that the operationalization of complexity as grade level confounds complexity per se with such related classificatory factors as the position's physical demands. Although these extraneous factors have a comparatively minor effect on grade determination, they could have nonetheless contaminated the measure of complexity used here. Although actual complexity scores were available for some occupations, using them would have required sacrificing the breadth and generalizability afforded by examining as many occupational units as were available.

Third, the forced categorization of tenure undoubtedly obscured and weakened its effect on agreement. As is well known (e.g., Pedhazur, 1997), categorizing an inherently continuous variable attenuates correlations. However, given that tenure had to be categorized to create groups of incumbents having a single score, the problem of categorization is not so much attenuation as it is the ability to form the appropriate groups. The relatively small number of new hires constrained the options available, with fewer than 12 months of experience the most appropriate definition of low tenure. Although the tenure groups used here are similar to those used elsewhere (Borman et al., 1996; Tross & Maurer, 2002), narrower groups defined in, say, one-month increments would have proved more enlightening. Such fine categorization, however, will be difficult to achieve given the typical distribution of tenure in an occupation.

Fourth, the conclusions that can be drawn from this study are limited by the absence of true-score estimates of KSAO importance. Knowing true scores would shed light on such issues as the extent to which agreement correlates with accuracy, the degree to which rating inflation artificially increases agreement, and the distribution of ratings around the true score (e.g., normal, positively skewed). Unfortunately, true scores were not available here, and could not be collected in any practical way. Job analysts could have rated the KSAOs (e.g., Jones et al., 2001; McCormick et al., 1977; Smith & Hakel, 1979), but it is unrealistic to assume that the analysts would be sufficiently familiar with each occupation and grade to provide passable true scores, nor is it realistic to assume that they would even be willing to rate so many occupational units. Although supervisory ratings are available for many of the series-grades examined here, there is no reason to believe that supervisory ratings are any closer to "true" than are incumbent ratings, and comparing the two would have gone beyond the intended scope of this study.

Fifth, the measurement of KSAO abstractness cannot be considered ideal. Each of two operationalizations were measured with single item ratings from only seven raters, who disagreed on nearly a quarter of the items. In addition, the raters could not be assembled to discuss their ratings, raising the possibility that different raters could have interpreted the same item in different ways. However, there is no assurance that increasing the number of items or raters would produce more stable or more accurate results. Although abstractness and multidimensionality are properties of the KSAOs, they do not—especially in the case of abstractness—exist outside the interpretation of the rater. Thus, applying an external standard, as that used here, will not produce as strong an effect as a wholly within-subjects design in which those who rate the importance of the KSAOs also rate the items.

The final limitation concerns the generalizability of the findings. Although the data include ratings from incumbents in a large number varied occupations—a total of 261 series-grades—they were all drawn from four broad occupational families (clerical, technical, professional, and administrative) in the Federal government. Given that public sector work is not inherently unique, these findings should, assuming similarly classified positions, generalize to the private sector. If, however, the classification system used here produces broader and more heterogeneous occupations than found in the private sector (or in other systems), then the present results may not fully generalize. Although an attempt was made to lessen aggregation, by examining agreement within agency and location, these factors are probably less important determinants of KSAO requirements than are task requirements. In other words, KSAO requirements are less affected by agency or location than by task requirements—controlling for agency and location cannot correct for an overly broad aggregation of positions based on their task requirements.

This present research has been concerned exclusively with ratings by incumbents, but supervisors and job analysts frequently provide KSAO ratings, as well. Because job analysts, supervisors, and incumbents have different perspectives on worker requirements (e.g., job analysts and supervisors know what *should be* needed, incumbents know what is *actually* needed), ratings need not converge across sources (Cornelius et al., 1984; Hazel et al., 1964; Huber, 1991; Hutt, 1996; Manson et al., 2000; Meyer, 1959; O'Reilly, 1973; Wilson, 1997). If job analysts and supervisors are more likely than incumbents to share the same prototype of the job, then job analysts and supervisors should agree more in their ratings. Research is generally consistent with this hypothesis (DeNisi et al., 1987; Friedman & Harvey, 1986; Hughes & Prien, 1989; Jones et al., 2001; McCormick et al., 1977; Smith & Hakel, 1979). Thus, one may expect agreement among job analysts and supervisors to be somewhat higher than the levels reported here.

Similarly, although the focus of the present research has been on importance ratings, other rating scales are also common (e.g., need for training, needed at entry, distinguishing value). One may hypothesize that the more a rating focuses on attributes of specific incumbents, the worse agreement will be. Two incumbents may, for example, agree on the importance of oral communication, but report quite different levels of training need. On the other hand, ratings such as importance and distinguishing value that focus on occupational characteristics should yield comparatively higher agreement. Thus, one would expect somewhat worse agreement on need for training scales and somewhat higher—though still low—agreement on needed at entry and distinguishing value scales.

Implications

Job Specification Validity

Validity can be defined as measuring what one intends to measure, i.e., the extent to which indicators of the construct of interest adequately represent that construct. From this construct-oriented definition, it follows that valid measures will have (a) content validity in so far as the indicators sample from the appropriate content universe, and (b) predictive validity in so far as the indicators, being representative of the underlying construct, relate to other constructs in the theoretically predicted manner. In the case of job specifications, where the underlying construct is usually the importance (or other characteristic) of a KSAO, content validity refers to the extent to which the rated KSAO encapsulates the entire construct domain, and predictive validity refers to the ability of the KSAO to predict job success.

Given the uses of job specifications (e.g., developing selection tests), practitioners are most concerned with their ability to identify truly important KSAOs, i.e., those that will predict job performance. Although, little direct evidence shows that SME ratings of importance correspond to their empirical importance as demonstrated by a criterion-related validity study (Tett, Holland, Hogan, & Burnett, 2002), the prevalent use and effectiveness of selection tests developed based on job specifications testifies to their validity. The issue, then, is not whether SMEs can identify important KSAOs, but rather the extent to which disagreement results in the exclusion of predictive KSAOs from job profiles.

As the results of this and other studies (e.g., Conte et al., 2003; Van Iddekinge, Putka, Raymark, & Eidson, 2003) show, confounding factors, such as demographic characteristics, job attitudes, and organizational performance, account for relatively little variance in ratings. The large proportion of unexplained variance most probably reflects the existence of meaningful subgroups within common job titles (e.g., Stutzman, 1983). Such is certainly the case in the Federal government, whose classification system must yield a manageable number of titles to cover two million employees working in more than 50 agencies. Indeed, the Office of Personnel Management acknowledges the problem of overly broad classifications, citing as an example the Human Resources Specialist (GS-201) occupation, "which resulted from the collapse of various HR positions into a single general classification. *While OPM's 201s share similar titles and classifications, they engage in work that emphasizes very different competencies*" (U.S. Office of Personnel Management, 2003, p. 13, emphasis added). The existence of such subgroups within a single job title seriously threatens the validity and usefulness of job specifications.

As shown in the supplemental analysis of dealing with disagreement, removing outliers to increase agreement has little effect on which KSAOs get classified as critical. One reason is that agreement and importance were positively correlated, decreasing the chance that removing outliers would change the mean importance of a KSAO. Another explanation is that the majority of ratings were distributed symmetrically about the mean. In any case, the analyses confirm that some KSAOs are important, on average, across many positions in an occupation. However, deleting raters to increase agreement overlooks KSAOs predictive of success in a subset of occupations. A selection battery developed based on the resulting job specification would tend to yield incumbents qualified to do a common set of tasks, but, potentially, not ideal for any position. Granted, such practices accomplish the objective of creating a mobile workforce, one capable of performing in any number of positions (e.g., Schippmann et al., 2002). But there is a point of diminishing returns: if classifications are too broad, and KSAO requirements too watered down, then, at the extreme, the probability that a new hire will possess the right KSAOs will not exceed chance levels, thereby defeating the purpose of the selection test and, by extension, the job specification as well. Thus, the true impact of disagreement in job specification ratings is not so much on predictive validity, but rather on utility, or what Sanchez and Levine (2000) call consequential validity.

In other words, a given KSAO that emerges as critical may very well be a strong predictor of job performance, yet the entire specification or job profile will not be content valid if other critical KSAOs are not identified. That is, a job specification will only be content valid if it identifies all the KSAOs predictive of success, and excludes those that are not (cf. Morgeson & Campion, 2000). The high levels of disagreement reported here and elsewhere suggest that job specifications may indeed suffer from poor content validity.

Most generally, disagreement implies that the ratings are not construct valid. If one makes the assumption that a true score exists, then ratings are valid only if they reflect that true score. Substantial disagreement reflects no common conceptualization of job requirements, and so cannot reflect a single true score. Note that the assumption of a true score cannot be avoided. Supervisors and job analysts, for example, rate the *job*. Although incumbents can be asked to rate the importance of a KSAO to their own *position* (e.g., if one believes that different profiles of KSAOs can be equally effective), at some point their ratings must be averaged to form a score for the entire *job*. Indeed, even asking the question “Is my job specification valid?” presupposes the existence of a true score, the construct being assessed. (Note that in the presence of multiple subgroups within a single job title, the putative true score has little meaning because it may not reflect the requirements of any of the groups.) Agreement by itself does not guarantee the validity of ratings, because for any number of reasons raters could agree on the wrong level; but there is no such thing as disagreeing on the right level. Thus, acceptable agreement is a necessary but not sufficient condition for construct validity. In light of the high levels of disagreement reported in this study, one must question the construct validity of job specifications.

Choice of Agreement Index

As the results show, different agreement indices can lead to different conclusions. For example, as seen in Hypothesis 1, different forms of r_{wg} led to drastically different estimates of agreement. Similarly, in the CT and PA occupations respectively, $r_{wg(EU)}$ and a_{wg} correlated r = .79 and r = .65, and led to the same conclusion about acceptable agreement 93% and 88% of the time. Thus, r_{wg} and a_{wg} share approximately 50% of their variance and lead to different conclusions about 10% of the time. However, the results of Hypotheses 2 through 4—although overall weak and inconsistent—were markedly different when using r_{wg} and a_{wg} as the criterion, perhaps primarily reflecting the strong relationship of r_{wg} with mean importance. The G-analyses were generally not consistent with the r_{wg} and a_{wg} analyses. Using r_{wg} and a_{wg}, agreement was higher in the PA than CT occupations, but unacceptable for both. On the other hand, the G-analyses showed that agreement was unacceptable for CT occupations but acceptable for PA occupations, a likely result of greater cross-item variance in the PA occupations. Overall, then, the level of assessed agreement depends on data characteristics and the index used.

The chosen index, therefore, must be appropriate to the analysis. In forming job specifications, the purpose of the analysis is to assess the extent to which a group of SMEs provide functionally interchangeable ratings. How well do r_{wg}, a_{wg}, and G-theory accomplish this objective?

The primary limitation of G-theory is its inability to assess agreement on a single item. As a consequence, it has little value in developing job specifications because single KSAOs, not multi-item scales, form the target of the analyses. Further, as seen here, if G-theory were used as an omnibus test of agreement, then substantial within-item variance can be masked by substantial between-item variance. For these reasons, G-theory should not be used in developing job specifications.

One limitation shared by r_{wg}, traditionally defined, and a_{wg} is their behavior at the small sample sizes characteristic of job specification studies. Because r_{wg} is a function of the ratio of observed to population variances, it will yield downwardly biased estimates of agreement. With large sample sizes, such as those used here, this bias will be small and inconsequential. But with more typical *n*s of 5 or 10, the obtained estimates can be misleading. To correct for this bias, researchers should use the sample rather than population variance. Although a_{wg} uses the sample variance, thereby yielding unbiased estimates at small sample sizes, it can produce a large number of uninterpretable values. For example, assuming a 1 – 5 scale, when $n = 5$ means less than 1.8 or greater than 4.2 are uninterpretable; when $n = 10$ these means shift to 1.4 and 4.6. As shown in the supplemental analyses, approximately 25% of a_{wg} values will be uninterpretable when $n = 5$, compared to 4% when $n = 10$. Therefore, to minimize the occurrence of uninterpretable values, a_{wg} should not be used with fewer than 10 raters.

The larger—and far more complex—issue concerns the choice of the reference variance when using r_{wg}. To simplify the discussion, a *fixed* variance will be defined as one derived from a distribution that is assumed to exist (e.g., the uniform distribution); a *floating* variance, on the other hand, is one derived from a distribution whose characteristics are at least in part determined by characteristics of the sample distribution (e.g., maximum possible variance at the observed mean). To appreciate the difference between fixed and floating variances, consider that the purpose of assessing agreement is to make an inference to the population of all possible raters (e.g., future incumbents, those not in the sample). As such, the reference variance should represent a realistic scenario of responding. In the absence of information about such a distribution, using information from the sample is the only way to estimate an appropriate reference variance.

The most common choice to date has been the variance of the uniform distribution. As a fixed variance, one must *assume* that random responding would yield rectangular ratings, with a mean of the scale midpoint. Any non-bimodal departure from the midpoint indicates some level of agreement, which, although presented as a weakness of this index, follows naturally from its conceptual underpinnings and makes perfect sense if, in fact, random ratings follow a uniform distribution. However, random ratings almost certainly do not follow such a distribution, and finding agreement leads to the untenable conclusion that the observed ratings are not representative of the population because they have different means. Therefore, r_{wg} using the uniform variance is conceptually inappropriate for job specifications.

More realistic floating reference variances can be estimated from characteristics of the sample. For example, Harvey and Hollander (2002) based r_{wg} on the uniform distribution bounded by the used range of observed responses. Such a strategy takes into account the likely range of random responding, and so is less lenient than basing the reference variance on the full scale. However, leniency can be replaced by severity if the used range yields a distribution similar to the distribution of responses. Basing the reference variance on a putative population distribution, as done here (see also Hollander & Harvey, 2003), can lead to a similar problem. If the population distribution reflects a high level of agreement, then little can be gained by introducing more agreement. For example, an observed variance of .20, indicating strong agreement, yields r_{wg} = .20 when compared to a variance of .25. The result is correct that .20 is a reduction of 20% from .25, but such a number fails to capture the degree to which the observed ratings are, in reality, quite interchangeable.

The a_{wg} index also uses a floating variance, requiring the minimal assumption that the mean of the reference distribution is the same as the observed mean. However, because a_{wg} uses the maximum possible variance at the mean, calculated as a function of the number of scale-low and scale-high responses, a_{wg} depends on the scale used. A given observed variance will yield a lower a_{wg} value if based on a 5-point scale than if based on a 10-point scale. Such a property requires the assumption that the difference between scale points is proportional to the number of scale points, that, for example, a variance of .25 indicates greater disagreement on a 5-point scale than on a 10-point scale. This assumption is reasonable if the 10-point scale is constructed to detect fine-grained differences; but if the scale is unrealistically broad (Harvey & Hollander, 2002), then a_{wg} will be upwardly biased. Another consequences of using the maximum possible variance at the observed mean is that the meaning of given variance depends on the mean. For example, a variance of .28 in a sample of 10 raters yields a_{wg} = .86 when M = 1.5 but .93 when M = 2.5. This result is consistent with the philosophy of a_{wg} (drawing randomly from a distribution with an extreme mean will result in higher agreement than drawing from a distribution with a moderate mean), but inconsistent with the goal of identifying interchangeability.

As the foregoing illustrates, all forms of r_{wg} and a_{wg} share the same shortcoming: they estimate the proportional reduction in variance rather than interchangeability per se. The question asked when developing a job specification is not the percentage reduction in variance, nor is it whether an estimated agreement level is different than chance (e.g., Cohen et al., 2001; Dunlap, Burke, & Smith-Crowe, 2003). The question is simply the extent to which different SMEs assign the same rating, within chance variation, to a KSAO. The different forms of r_{wg} and a_{wg}, in addition to the average deviation index (Burke et al., 1999) and even the standard deviation, are all based on the same fundamental unit, the variance. The indices simply differ in what "chance" variation looks like. As a result, all forms will be highly correlated. The irreducible problem, then, is finding a criterion that indicates acceptable agreement. Thus, researchers should choose an index that conforms to their data and assumptions, and empirically or rationally justify (1) their choice of index and (2) their choice for a criterion of agreement.

Implications for Practice

The results of this study show that incumbents disagree substantially in their ratings of KSAO importance. In general, the broader the classification system, the more disagreement researchers can expect. Based on this and prior research, the disagreement seems to stem from real cross-position and subgroup differences. Given the purpose of job specifications—to describe job-level requirements—the challenge researchers and practitioners must face is how to obtain the most useful and predictive information within the constraints of a manageable number of job titles.

The governmentwide studies from which this study's data were taken grew out of a desire to create a central clearinghouse for job analytic data. Rather than forcing agencies to conduct their own job analyses and specifications, which would result in the duplication of efforts, the agencies can draw on a single source of data. Creating such a database, however, required the development of a single task and KSAO list, to the exclusion of occupational specific KSAOs. The PA study, for example, shows that physical strength is not important to psychologists, but says nothing about the importance of research skills. Likewise, little information exists to differentiate between clinical and industrial and organizational psychologists. Combined with a broad classification system that, for example, does not distinguish between types of psychology, these studies produce only marginally useful information. Anecdotal reports and personal experience show that the data from these studies must always be augmented with additional data collections in order to produce the type and quality of data needed to develop selection systems.

The O*NET grew out of a similar desire to create a clearinghouse for job analytic data, and suffers from the same problems. Broad classification, coupled with questionable measurement practices, has produced substantial disagreement in ratings (Harvey & Hollander, 2003; Hollander & Harvey, 2003). As a result, the O*NET may not fulfill the objectives for which it was developed. For example, the O*NET may not provide enough information on which to base decisions of social security disability compensation (Wilson, 2003).

However, data such as those just described are not completely without merit. Broad information will suffice for some uses, such as vocational guidance, and can be used to some extent in practical applications, such as job component validation (e.g., Jeanneret & Strong, 2003). In general, though, the usefulness of job specification data decreases as classification breadth and disagreement increase. If a job specification is too broad and contains too much disagreement, then the results will apply to few if any positions within the title. On the other hand, if a specification is too narrow, then it will apply to only one or a few positions, thereby limiting its usefulness. Researchers must rationally choose an approach to job specification that yields data appropriate to the intended applications: selection and training, for instance, will require greater detail, vocational guidance somewhat less.

Because job specifications are seldom conducted for a single purpose, researchers should choose to error on the side of providing too much detail. Low-level information can always be aggregated upward, but high-level data cannot be disaggregated to lower levels. Although collecting specific information may not always be practical with traditional paper-and-pencil methods, computer adaptive KSAO lists could be constructed in such a way that SMEs rate only relevant KSAOs (cf. the Common Metric Questionnaire; Harvey, 1993). Such technology removes many of the practical barriers to collecting narrow and focused job information.

The unfortunate reality is that users of the O*NET and similar data will likely encounter substantial disagreement in KSAO ratings. If the disagreement stems from the inappropriate aggregation of distinct job units, then, as mentioned above, the only remedy will often be to collect the additional information necessary to construct a meaningful and useful occupational profile. Researchers should not treat disagreement as error, but rather as an indication that the current method fails to adequately capture distinctions among individual and groups of positions. That is, researchers should not treat disagreement as nuisance variance, but rather as an opportunity to better understand the nature of the work and its requirements.

Future Directions

Although much of the discussion has centered on the existence of meaningful subgroups of positions within a single occupational title, such a proposition rests mainly on empirical work showing that potential moderators of agreement account for little variance in ratings. Relatively little research has directly explored the existence of within-title subgroups (Green & Stutzman, 1986; Harvey, 1986; Schmitt & Cohen, 1989; Stutzman 1983). To extend this research, one could first obtain validated position-level task ratings and ratings of KSAO importance, and then aggregate the positions to various extents to determine at what point agreement becomes unacceptable. Such information could be used to better understand the source of the disagreement reported here, and to guide future classification efforts.

Another line of research could explore the practical consequences of inappropriately broad classifications. Discussions of this issue have to date been largely anecdotal or theoretical, focusing on such issues as the aggregation bias (James, 1982). Research is needed to show the extent to which the content validity of job specifications does indeed suffer when broad classifications fail to make fine distinctions among subgroups of positions.

Other research could refine and extend the search for moderators of agreement. For example, no attempt was made here to link task ratings to KSAO ratings. Such an analysis, though beyond the scope of this study, could shed light on the degree to which disagreement in KSAOs reflects variance in task requirements. The role of experience could also be explicated with a longitudinal study of how work roles develop over time, when different positions diverge or converge, and if experienced incumbents really do, as originally hypothesized here, gain a perspective on the entire job. Other research could further explore the role of complexity using more tightly defined occupational groupings, perhaps based on work roles such as supervisor or project manager. Sharing a similar role within a single title should result in greater agreement on role-related KSAOs, regardless of occupational complexity. Researchers should also continue to search for item characteristics that influence agreement. Better controlled studies, for example, could elucidate the effects of abstractness and multidimensionality. In short, research is needed to account for the large proportion of unexplained variance in ratings.

Future research should also refine the measurement of agreement. Researchers could, for example, explore whether the level of acceptable disagreement varies for different applications (e.g., selection vs. vocational guidance). In light of the limitations or r_{wg}, a_{wg}, and G-theory, additional research is also needed on alternative measures of agreement. That is, none of the indices reviewed here quite captures the most vital and useful information, viz., the extent to which the observed variance is within the range of what would be expected by change. One alternative to r_{wg} and a_{wg} may be an index that gives the probability that a randomly chosen rating will be within the 95% CI of the mean or median rating. Such an index would give a direct estimate of interchangeability.

Conclusion

This study represents one of the most comprehensive to date, measuring agreement, using multiple methods, among more than 38,000 incumbents in 61 occupational series and 261 series-grades. As predicted, agreement failed to reach acceptable levels in nearly every case. However, contrary to expectation, experience, occupational complexity, and KSAO abstractness accounted for little of the disagreement. Although the reasons for these null findings are not entirely clear, the most likely explanation is that true cross-position variance simply overshadowed the variance due to these rater, occupation, and item characteristics. If so, then the disagreement reported here reflects a coarse classification system that inadequately distinguishes among meaningful subgroups within single occupational titles. The existence of such subgroups threatens not so much the predictive validity of job specifications as their content validity, in so far as predictive KSAOs are not identified as such. Future research must focus on the existence of such subgroups, their consequences, and ways of identifying them.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Albemarle Paper Co. v. Moody (1975). 422 US 405.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. APA, Washington DC.

Arvey, R. D., Davis, G. A., McGowen, S. L., & Dipboye, R. L. (1982). Potential sources of bias in job analytic processes. *Academy of Management Journal, 25*, 618-629.

Ash, R. A., & Edgell, S. L. (1975). A note on the readability of the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology, 60*, 765-766.

Berry, K. J., & Mielke, P. W., Jr. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement, 48*, 921-933.

Borman, W. C., Dorsey, D., & Ackerman, L. (1992). Time-spent responses as time allocation strategies: Relations with sales performance in a stockbroker sample. *Personnel Psychology, 45*, 763-777.

Brennan, R. L., & Krane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14, 277-289.

Brown, R. D., & Hauenstein, N. M. (2003). *Interrater agreement reconsidered: An alternative to the* r_{wg} *indices*. Manuscript submitted for publication.

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5*, 159-172.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2*, 49-68.

Burnkrant, S. R. (2001, April). *Effects of Competition on faking job-specific profiles: Job-desirability or social desirability?* Paper presented at the 16 Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Butler, S. K., & Harvey, R. J. (1988). A comparison of holistic versus decomposed rating of Position Analysis Questionnaire work dimensions. *Personnel Psychology, 41*, 761-771.

Cain, P. S., & Green, B. F. (1983). Reliabilities of selected ratings available from the Dictionary of Occupational Titles. *Journal of Applied Psychology, 68*, 155-165.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement, 13*, 119-135.

Cascio, W. F. (1998). *Applied psychology in human resource management* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New-York : John Wiley.

Church, K., & Sher, M. (1998). *Dimension of effective behavior: Professional and administrative occupations*. (Rep. No. PRDC-98-05). Washington, D.C.: U.S. Office of Personnel Management, Personnel Resources and Development Center.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37- 46.

Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the $r_{wg(j)}$ index of agreement.

Psychological Methods, 6, 297-310.

Conley, P. R., & Sackett, P. R. (1987). Effects of using high- versus low-performing job incumbents as sources of job analysis information. *Journal of Applied Psychology, 72*, 434-437.

Conte, J. M., Dean, M. A., Ringenbach, K. L., Moran, S. K., & Landy, F. J. (2003). *The association of work attitudes with job analysis ratings*. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando.

Contreras v. City of Los Angeles (1981). 25 FEP 867.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, and self-ratings. *Human Performance, 10*, 331-360.

Cornelius, E. T., DeNisi, A. S., & Blencoe, A. G. (1984). Expert and naïve raters using the PAQ: Does it matter? *Personnel Psychology, 37*, 453-464.

Cornelius, E. T., III, & Lyness, K. S. (1980). A comparison of holistic and decomposed judgment strategies in job analyses by job incumbents. *Journal of Applied Psychology, 65*, 155-163.

Curnow, C., McGonigle, T., & Sideman, L. (2003, April). Comparing reliability and agreement on O*Net vs. job-specific job analysis questionnaires. In M. Buster (Chair), *Addressing some limitations in the science and practice of job analysis*. Symposium presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Dansereau, F., Alutto, J. A., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The variant approach*. Englewood Cliffs, NJ: Prentice Hall.

DeNisi, A. S., & Shaw, J. B. (1977). Investigation of the uses of self-reports of abilities. *Journal of Applied Psychology, 62*, 641-644.

DeNisi, A. S., Cornelius, E. T., & Blencoe, A. G. (1987). Further investigation of common knowledge effects on job analysis ratings. *Journal of Applied Psychology, 72*, 262-268.

DeNisi, A. S., & Williams, K. J. (1988). Cognitive approaches to performance appraisal. *Research in Personnel and Human Resource Management, 6*, 109-155.

Dierdorff, E. C., & Wilson, M. A. (2003). A meta-analysis of job analysis reliability. *Journal of Applied Psychology, 88*, 635-646.

Doverspike, D., Carlisi, A. M., Barrett, G. V., & Alexander, R. A. (1983). Generalizability analysis of a point-method job evaluation instrument. *Journal of Applied Psychology, 68*, 476-483.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for r_{wg} and average deviation interrater agreement indexes. *Journal of Applied Psychology, 88*, 356-362.

EEOC v. Atlas Paper Box Co. (1989). 868 F.2d 1487.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive process in performance appraisal. *Journal of Applied Psychology, 66*, 127-148.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30*, 71-76.

Fleishman, E. A., & Mumford, M. D. (1991). Evaluating classifications of job behavior: A construct validation of the ability requirements scales. *Personnel Psychology, 44*, 523-575.

Friedman, L., & Harvey, R. J. (1986). Can raters with reduced job descriptive information provide accurate Position Analysis Questionnaire (PAQ) ratings? *Personnel Psychology, 39*, 779-790.

Geyer, P. D., Hice, J., Hawk, J., Boese, R. & Brannon, Y. (1989). Reliabilities of ratings available from the Dictionary of Occupational Titles. *Personnel Psychology, 42*, 547-560.

Graen, G. (1976). Role-making processes within complex organizations. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology*. Chicago: Rand McNally.

Graen, G. B., & Scandura, T. A. (1987). Toward a psychology of dyadic organizing. *Research in Organizational Behavior, 9*, 175-208.

Green, S. B., & Stutzman, T. M. (1986). An evaluation of methods to select respondents to structured job-analysis questionnaires. *Personnel Psychology, 39*, 543-564.

Green, S. B., & Veres, J. G. (1990). Evaluation of an index to detect inaccurate respondents to a task analysis inventory. *Journal of Business and Psychology, 5*, 47-61.

Griggs v. Duke Power Co. (1971). 401 US 424.

Guardians Association of the New York City Police Department v. Civil Service Commission of the City of New York (1980). 23 FEP 909.

Hahn, D. C., & Dipboye, R. L. (1988). Effects of training and information on the accuracy and reliability of job evaluations. *Journal of Applied Psychology, 73*, 146-153.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*, 43-62.

Harris, M. (1998). Competency modeling: Viagraized job analysis or impotent imposter? *The Industrial and Organizational Psychologist, 36.*

Harvey, R. J. (1986). Quantitative approaches to job classification: A review and critique. *Personnel Psychology, 39*, 267-289.

Harvey, R. J. (1991). Job analysis. In M. Dunnette & L. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.

Harvey, R. J. (1993). *Research monograph: The development of the CMQ*. San Antonio: The Psychological Corporation.

Harvey, R. J., & Hayes, T. L. (1986). Monte Carlo baselines for interrater reliability correlations using the Position Analysis Questionnaire. *Personnel Psychology, 39,* 345-357.

Harvey, R. J., & Hollander, E. (2002, April). Assessing interrater agreement in the O*NET. In M. A. Wilson (Chair), *The O*NET: Mend it or end it?* Symposium presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto.

Harvey R. J. & Lozada-Larsen, S. R. (1988). Influence of amount of job descriptive information on job analysis rating accuracy. *Journal of Applied Psychology, 73*, 457-461.

Harvey, R. J., & Wilson, M. A. (1998, April*)*. Monte Carlo baselines for interrater agreement when rating KSA requirements: How much is enough? In R. J. Harvey (Chair), *Measurement issues in job analysis: Good news and bad news*. Symposium presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis.

Harvey, R. J., & Wilson, M. A. (2000). Yes Virginia, there *is* an objective reality in job analysis. *Journal of Organizational Behavior, 21*, 829-854.

Harvey, R. J., Wilson, M. A., & Blunt, J. H. (1994, April). *A comparison of rational/holistic versus empirical/decomposed methods of identifying and rating general work behaviors*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Nashville.

Hazel, J. T., Madden, J. M., & Christal, R. E. (1964). Agreement between work-supervisor descriptions of the worker's job. *Journal of Industrial Psychology, 2*, 71-79.

Hollander, E., & Harvey, R. J. (2002, April). Generalizability theory analysis of item-level O*NET database ratings. In M. A. Wilson (Chair), *The O*NET: Mend it or end it?* Symposium presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Huber, V. L. (1991). Comparison of supervisor-incumbent and female-male multidimensional job evaluation ratings. *Journal of Applied Psychology, 76*, 115-121.

Hughes, G. L., & Prien, E. P. (1989). Evaluation of task and job skill linkage judgments used to develop test specifications. *Personnel Psychology, 42*, 283-292.

Hutt (1996).

Ilgen, D. R., & Hollenbeck, J. R. (1991). The structure of work: Job design and roles. In M. D. Dunnette, & L. M. Hough (Eds), *Handbook of industrial and organizational psychology*, Vol. 2 (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67*, 219-229.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*, 85-98.

James, L. R., Demaree, R. G., & Wolf, G. (1993). r_{wg}: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78*, 306-309.

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement, 61*, 277-289.

Jeanneret, P. R., Borman, W. C., Kubisiak, U. C., & Hanson, M. A. (1999). Generalized work activities. In N. G. Peterson & M. D. Mumford (eds.), *An occupational information system for the 21st century: The development of the O*NET.*

Jeanneret, P. R., Strong, M. H. (2003). Linking O*Net job analysis information to job requirement predictors: An O*Net application. *Personnel Psychology, 56*, 465-492.

Jones v. New York City Human Resources Administration (1975). 12 FEP 265.

Jones, A. P., Main, D. S., Butler, M. C., & Johnson, L. A. (1982). Narrative job descriptions as potential sources of job analysis ratings. *Personnel Psychology, 35*, 813-828.

Jones, R. G., Sanchez, J. I., Parameswaran, G., Phelps, J., Shoptaugh, C., Williams, M., & White, S. (2001). Selection or training? A two-fold test of the validity of job-analytic ratings of trainability. *Journal of Business and Psychology, 15*, 363-389.

Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoek, J. D., & Rosenthal, R. A. (1964). *Organizational stress: Studies in role conflict and role ambiguity*. New York: Wiley.

Kerber, K. W., & Campbell, J. P. (1987). Correlates of objective performance among computer salespeople: Tenure, work activities, and turnover. *Journal of Personnel Selling and Sales Management, 7*, 39-50.

Kesselman, G. A., & Lopez, F. E. (1979). The impact of job analysis on employment test validation for minority and nonminority accounting personnel. *Personnel Psychology, 32*, 91-108.

Kozlowski, S. W., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77*, 161-167.

Landy, F. J., & Vasey, J. (1991). Job analysis: The composition of SME samples. *Personnel*

Psychology, 44, 27-50.

Lawshe, C. H. (1975). *A quantitative approach to content validity. Personnel Psychology, 28*, 563-575.

Lindell, M. K. (2001). Assessing and testing interrater agreement on a single target using multi-item rating scales. *Applied Psychological Measurement, 25*, 89-99.

Lindell, M. K., & Brandt, C. J. (1997). Measuring interrater agreement for ratings of a single target. *Applied Psychological Measurement, 21*, 271-278.

Lindell, M. K., & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of CVI, T, $r_{wg(j)}$ and $r^*_{wg(j)}$ indexes. *Journal of Applied Psychology, 85*, 331-348.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement, 23*, 127-135.

Lindell, M. K., Clause, C. S., Brandt, C. J., & Landis, R. S. (1998). Relationship between organizational context and job analysis task ratings. *Journal of Applied Psychology, 83*, 769-776.

Lopez, F. M., Kesselman, G. A., & Lopez, F. E. (1981). An empirical test of a trait-oriented job analysis technique. *Personnel Psychology, 34*, 479-502.

Love, K. G., Bishop, R. C., & Scionti, C. (1991). Response bias in job analysis ratings: Relation between ratings of task liking and task characteristics. *Psychological Reports, 68*, 1113-1114.

Manson, T. M., Levine, E. L., & Brannick, M. T. (2000). The construct validity of task inventory ratings: A multitrait-multimethod analysis. *Human Performance, 13*, 1-22.

Maxwell, A, E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry, 130*, 79-83.

McCormick, E. J., DeNisi, A. S., & Shaw, J. B. (1979). Use of the Position Analysis Questionnaire for establishing the job component validity of tests. *Journal of Applied Psychology, 64*, 51-56.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of the job characteristics and job dimensions as based on the Position Analysis Questionnaire. *Journal of Applied Psychology, 56*, 347-368.

McCormick, E. J., Mecham, R. C., & Jeanneret, P. R. (1977). *Technical manual for the Position Analysis Questionnaire (PAQ) (System II)*. West Lafayette, IN: University Book Store.

McGraw, K. O, & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

Meyer, H. H. (1959). A comparison of foreman and general foreman conceptions of the foreman's job responsibilities. *Personnel Psychology, 12*, 445-452.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 82*, 627-655.

Morgeson, F. P., & Campion, M. A. (2000). Accuracy in job analysis: Toward an inference-based model. *Journal of Organizational Behavior, 21*, 819-827.

Mullins, W. C., & Kimbrough, W. W. (1988). Group composition as a determinant of job analysis outcomes. *Journal of Applied Psychology, 73*, 657-664.

Murphy, K. R., & DeShon, R. (2000a). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873-900.

Murphy, K. R., & DeShon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology, 53*, 913-924.

Murphy, K. F., & Wilson, M. A. (1997, April). *Estimating the reliability of job analysis ratings: The role of level of abstraction, method of estimation and modality*. Paper presented at the 12th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology, 75*, 77-86.

O'Reilly, A. P. (1973). Skill requirements: Supervisor-subordinate conflict. *Personnel Psychology, 26*, 75-80.

Oshagbemi, T. (1999). Overall job satisfaction: How good are single versus multiple-item measures? *Journal of Managerial Psychology, 14*, 388-403.

Paunonen, S V., & Jackson, D. N. (1987). Accuracy of interviewers and students in identifying the personality characteristics of personnel managers and computer programmers. *Journal of Vocational Behavior, 31*, 26-36.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth: Harcourt Brace.

Pine, D. E. (1995). Assessing the validity of job ratings: An empirical study of false reporting in task inventories. *Public Personnel Management, 24*, 451-460.

Pollack, L., Simons, C., Patel, R., & Gregory, D. (2000). *Federal professional and administrative occupations: An application of the Multipurpose Occupational Systems Analysis Inventory—Closed-Ended (MOSAIC)* (Rep. No. PRDC-0399). Washington, D.C.: U.S. Office of Personnel Management, Personnel Resources and Development Center.

Pulakos, E. D., & Wexley, K. N. (1983). The relationship among perceptual similarity, sex, and performance ratings in manager-subordinate dyads. *Academy of Management Journal, 26*, 129-139.

Reed, P., & Jackson, D. N. (1975). Clinical judgment of psychopathology: A model for inferential accuracy. *Journal of Abnormal Psychology, 84*, 475-482.

Richman, W. L., & Quinones, M. A. (1996). Task frequency rating accuracy: The effect of task engagement and experience. *Journal of Applied Psychology, 81*, 512-524.

Rodriguez, D. A., Usala, P., & Shoun, S. (1996). *Federal clerical and technical occupations: An application of the Multipurpose Occupational Systems Analysis Inventory—Closed Ended (MOSAIC)* (Rep. No. PRDC-95-08). Washington, D.C.: U.S. Office of Personnel Management, Personnel Resources and Development Center.

Rothstein, H. R. (1990) Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75*, 322-327

Rothstein, M., & Jackson, D. N. (1980). Decision making in the employment interview: An experimental approach. *Journal of Applied Psychology, 65*, 271-283.

Sanchez, J. I., & Fraser, S. L. (1992). On the choice of scales for task analysis. *Journal of Applied Psychology, 77*, 545-553.

Sanchez, J. I., & Levine, E. L. (1994). The impact of raters' cognition on judgment accuracy: An extension to the job analysis domain. *Journal of Business and Psychology, 9*, 47-57.

Sanches, J. I., & Levine, E. L. (2000). Accuracy or consequential validity: Which is the better standard for job analysis data? *Journal of Organizational Behavior, 21*, 809-818.

Schay, B. W., Buckley, T., Chmielewski, M., Medley-Proctor, K., & Burnkrant, S. R. (2001). *Validating the use of fewer-than-nine factors in a leveling tool*. Washington, D.C.: U.S.

Office of Personnel Management, Personnel Resources and Development Center.

Schippmann, J. S., Ash, R. A., Battista, M., Carr, L., Eyde, L. D., Hesketh, B., Kehoe, J., Pearlman, K., Prien, E. P., & Sanchez, J. I. (2000). The Practice of Competency Modeling. *Personnel Psychology, 53*, 703-740.

Schmitt, N., & Cohen, S. A. (1989). Internal analysis of task ratings by job incumbents. *Journal of Applied Psychology, 74*, 96-104.

Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology, 74*, 368-370.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. Personnel Psychology, 53, 901-912.

Schneider, W., Dumais, S. T., & Shiffrin, R. M. (1984). Automatic and control processing and attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 1-27). Orlando, FL: Academic Press.

Schneider, W., Dumais, S. T., & Shiffrin, R. M. (1984). Automatic and control processing and attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 1-27). Orlando, FL: Academic Press.

Schuster, C., & Smith, D. A. (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Methods, 7*, 384-395.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oak, CA: Sage.

Shoun, S. (1995). *Dimensions of effective behavior: Clerical and technical employees*. (Rep. No. PRDC-95-02). Washington, D.C.: U.S. Office of Personnel Management, Personnel

Resources and Development Center.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Silverman, S. B., Wexley, K. N., & Johnson, J. C. (1984). The effects of age and job experience on employee responses to a structured job analysis questionnaire. *Public Personnel Management, 13*, 355-359.

Sledge v. J. P. Stevens & Co. (1978). 585 F.2d 625.

Smith, J. E., & Hakel, M. D. (1979). Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. *Personnel Psychology, 32*, 677-692.

Society for Industrial and Organizational Psychology (1987). *Principles for the validation and use of personnel selection procedures*.

Spector, P. E. (2000). *Industrial and organizational psychology: Research and practice* (2nd ed.). New York: Wiley.

Stutzman, T. M. (1983). Within classification job differences. *Personnel Psychology, 36*, 503-516.

Surrette, M. A., Aamodt, M. G., & Johnson, D. L. (1990). Effects of analyst training and amount of available job related information on job analysis ratings. *Journal of Business and Psychology, 4*, 439-451.

Tajfel, H., & Turner, J.C. (1986). The social identity theory of intergroup behavior. In S.Worchel & W.G.Austin. (Eds.). *Psychology of Intergroup Relations*, pp. 7-24. Chicago: Nelson.

Tett, R. P., Holland, B., Hogan, J., & Burnett, D. D. (2002, April). *Validity of trait-based job analysis using moderator correlations*. Paper presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22*, 358-376.

Tross, S. A., & Maurer, T. J. (2000). The relationship between SME job experience and job analysis ratings: Findings with and without statistical control. *Journal of Business and Psychology, 15*, 97-110.

Tross, S. A., & Maurer, T. J. (2002). Supervisors as SMEs: The relationship between supervisor job experience and ratings of subordinate skill importance. *Journal of Business and Psychology, 16*, 413-430.

United States Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, Department of Justice. (1978). *Uniform guidelines on employee selection procedure.* Federal Register, 43, (166), 38290-38315.

United States Office of Personnel Management (2003). *The plan for the strategic management of human capital*. Draft strategic plan, Washington, D.C.

Van Iddekinge, C. H., Putka, D. J., Raymark, P. H., & Eidson, C. E., Jr. (2003, April). *Sources of variance in worker-oriented job analysis ratings*. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando.

Veres, J. G., Green, S. B., & Boyles, W. R. (1991). Racial differences on job analysis questionnaires: An empirical study. *Public Personnel Management, 20*, 135-144.

Vulcan Society v. Civil Service Commission (1973). 490 F.2d 387.

Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods, 4*, 361-375.

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology, 82*, 247-252.

Wexley, K. N., & Silverman, S. B. (1978). An examination of differences between managerial effectiveness and response patterns on a structured job analysis questionnaire. *Journal of Applied Psychology, 63*, 646-649.

Wilson, M. A. (1997). The validity of task coverage ratings by incumbents and supervisors: Bad news. *Journal of Business and Psychology, 12*, 85-95.

Wilson, M. A. (2002, April). *The O*NET: Mend it or end it?* Symposium presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto.

Wright, P. M., Anderson, C., Tolzman, K., & Helton, T. (1990). An examination of the relationship between employee performance and job analysis ratings. *Academy of Management Best Papers*, 50, 299-303.

Appendix A:

Clerical and Technical KSAOs

1. *Reading* - Learns from written material by determining the main idea or essential message. Recognizes correct English grammar, punctuation, and spelling
2. *Writing* - Uses correct English grammar, punctuation, and spelling to communicate thoughts, ideas, information, and messages in writing
3. *Listening* - Receives, attends to, interprets, and responds to verbal messages and other cues such as body language in ways that are appropriate to listeners and situations
4. *Speaking* - Uses correct English grammar to organize and communicate ideas in words that are appropriate to listeners and situations; uses body language appropriately
5. *Arithmetic/Mathematical Reasoning* - Performs computations such as addition, subtraction, multiplication, and division correctly; solves practical problems by choosing appropriately from a variety of mathematical technique.
6. *Reasoning* - Discovers or selects rules, principles, or relationships between facts and other information
7. *Decision Making* - Specifies goals and obstacles to achieving those goals, generates alternatives, considers risks, and evaluates and chooses the best alternative, etc.
8. *Creative Thinking* - Uses imagination to combine ideas or information in new ways
9. *Mental Visualization* - Sees things in the mind by mentally organizing and processing symbols, pictures, graphs, objects, or other information
10. *Memory* - Recalls information that has been presented previously
11. *Eye-Hand Coordination* - Accurately coordinates one's eyes with one's fingers, wrist, or arms to move, carry, or manipulate objects, or to perform other job-related tasks

12. *Perceptual Speed* - Sees detail in words, numbers, pictures, and graphs, quickly and accurately
13. *Physical Strength and Agility* - Ability to bend, lift, climb, stand, and walk for long periods of time; ability to perform moderately heavy laboring work
14. *Stamina* - Performs repetitive tasks effectively over a long period of time, for example, data entry and coding
15. *Applies Technology to Tasks* - Selects and understands procedures, machines, or tools that will produce the desired results; identifies or solves problems in machines, computers, etc.
16. *Technical Competence* - Knowledge of how to perform one's job. Refers to specialized knowledge that is acquired through formal training or extensive on-the-job experience
17. *Organizational Awareness* - Knows how social, political, organizational, and technological systems work and operates effectively within them
18. *Manages and Organizes Information* - Identifies a need; gathers, organizes, and maintains information; determines its importance and accuracy, and communicates it by a variety of methods
19. *Manages Resources* - Selects, acquires, stores, and distributes resources such as materials, equipment, or money
20. *Manages Human Resources* - Plans, distributes, and monitors work assignments; evaluates work performance and provides feedback to others on their performance
21. *Conscientiousness* - Displays a high level of effort and commitment towards performing work; demonstrates responsible behavior
22. *Integrity/Honesty* - Displays high standards of ethical conduct and understands the impact of violating these standards on an organization, self, and others; chooses an ethical course of

action, etc.

23. *Interpersonal Skills* - Shows understanding, friendliness, courtesy, tact, empathy, cooperation, concern, and politeness to others; relates well to different people from varied backgrounds and different situations

24. *Self-Esteem* - Believes in own self-worth, maintains a positive view of self, and displays a professional image

25. *Self-Management* - Sets well-defined and realistic personal goals; monitors progress and is motivated to achieve; manages own time and deals with stress effectively

26. *Flexibility* - Adapts quickly to changes

27. *Leadership* - Interacts with others to influence, motivate, and challenge them

28. *Teaches Others* - Helps others learn; identifies training needs; provides constructive reinforcement; coaches others on how to perform tasks; acts as a mentor

29. *Teamwork* - Encourages and facilitates cooperation, pride, trust, and group identity; fosters commitment and team spirit; works with others to achieve goals

30. *Negotiation* - Works with others towards an agreement that may involve exchanging specific resources or resolving differences

31. *Customer Service* - Works and communicates with clients and customers to satisfy their expectations. Committed to quality services

Appendix B:

Professional and Administrative KSAOs

1. *Reading* - Understands and interprets written material, including technical material, rules, regulations, instructions, reports, charts, graphs, or tables; applies what is learned from written material to specific situations.
2. *Writing* - Recognizes or uses correct English grammar, punctuation, and spelling; communicates information (for example, facts, ideas, or messages) in a succinct and organized manner; produces written information, which may include technical material, that is appropriate for the intended audience.
3. *Arithmetic* - Performs computations such as addition, subtraction, multiplication, and division correctly using whole numbers, fractions, decimals, and percentages.
4. *Mathematical Reasoning* - Solves practical problems by choosing appropriately from a variety of mathematical and statistical techniques.
5. *Oral Communication* - Expresses information (for example, ideas or facts) to individuals or groups effectively, taking into account the audience and nature of the information (for example, technical, sensitive, controversial); makes clear and convincing oral presentations; listens to others, attends to nonverbal cues, and responds appropriately.
6. *Creative Thinking* - Uses imagination to develop new insights into situations and applies innovative solutions to problems; designs new methods where established methods and procedures are inapplicable or are unavailable.
7. *Information Management* - Identifies a need for and knows where or how to gather information; organizes and maintains information or information management systems.
8. *Decision Making* - Makes sound, well-informed, and objective decisions; perceives the

impact and implications of decisions; commits to action, even in uncertain situations, to accomplish organizational goals; causes change.

9. *Reasoning* - Identifies rules, principles, or relationships that explain facts, data, or other information; analyzes information and makes correct inferences or draws accurate conclusions.
10. *Problem Solving* - Identifies problems; determines accuracy and relevance of information; uses sound judgment to generate and evaluate alternatives, and to make recommendations.
11. *Mental Visualization* - Sees things in the mind by mentally organizing and processing symbols, pictures, graphs, objects, or other information (for example, sees a building from a blueprint, or sees the flow of work activities from reading a work plan).
12. *Learning* - Uses efficient learning techniques to acquire and apply new knowledge and skills; uses training, feedback, or other opportunities for self-learning and development.
13. *Self-Esteem* - Believes in own self-worth; maintains a positive view of self and displays a professional image.
14. *Teamwork* - Encourages and facilitates cooperation, pride, trust, and group identity; fosters commitment and team spirit; works with others to achieve goals.
15. *Integrity/Honesty* - Contributes to maintaining the integrity of the organization; displays high standards of ethical conduct and understands the impact of violating these standards on an organization, self, and others; is trustworthy.
16. *Self-Management* - Sets well-defined and realistic personal goals; displays a high level of initiative, effort, and commitment towards completing assignments in a timely manner; works with minimal supervision; is motivated to achieve; demonstrates responsible behavior.
17. *Interpersonal Skills* - Shows understanding, friendliness, courtesy, tact, empathy, concern,

and politeness to others; develops and maintains effective relationships with others; may include effectively dealing with individuals who are difficult, hostile, or distressed; relates well to people from varied backgrounds and different situations; is sensitive to cultural diversity, race, gender, disabilities, and other individual differences.

18. *Planning and Evaluating* - Organizes work, sets priorities, and determines resource requirements; determines short- or long-term goals and strategies to achieve them; coordinates with other organizations or parts of the organization to accomplish goals; monitors progress and evaluates outcomes.
19. *Attention To Detail* - Is thorough when performing work and conscientious about attending to detail.
20. *Financial Management* - Prepares, justifies, and/or administers the budget for program areas; plans, administers, and monitors expenditures to ensure cost-effective support of programs and policies; assesses financial condition of an organization.
21. *Managing Human Resources* - Plans, distributes, coordinates, and monitors work assignments of others; evaluates work performance and provides feedback to others on their performance; ensures that staff are appropriately selected, utilized, and developed, and that they are treated in a fair and equitable manner.
22. *Leadership* - Influences, motivates, and challenges others; adapts leadership styles to a variety of situations.
23. *Teaching Others* - Helps others learn through formal or informal methods; identifies training needs; provides constructive feedback; coaches others on how to perform tasks; acts as a mentor.
24. *Customer Service* - Works with clients and customers (that is, any individuals who use or

receive the services or products that your work unit produces, including the general public, individuals who work in the agency, other agencies, or organizations outside the Government) to assess their needs, provide information or assistance, resolve their problems, or satisfy their expectations; knows about available products and services; is committed to providing quality products and services.

25. *Organizational Awareness* - Knows the organization's mission and functions, and how its social, political, and technological systems work and operates effectively within them; this includes the programs, policies, procedures, rules, and regulations of the organization.
26. *External Awareness* - Identifies and understands economic, political, and social trends that affect the organization.
27. *Vision* - Understands where the organization is headed and how to make a contribution; takes a long-term view and recognizes opportunities to help the organization accomplish its objectives or move toward the vision.
28. *Influencing/Negotiating* - Persuades others to accept recommendations, cooperate, or change their behavior; works with others towards an agreement; negotiates to find mutually acceptable solutions.
29. *Conflict Management* - Manages and resolves conflicts, grievances, confrontations, or disagreements in a constructive manner to minimize negative personal impact.
30. *Stress Tolerance* - Deals calmly and effectively with high stress situations (for example, tight deadlines, hostile individuals, emergency situations, dangerous situations).
31. *Flexibility* - Is open to change and new information; adapts behavior or work methods in response to new information, changing conditions, or unexpected obstacles; effectively deals with ambiguity.

32. *Technology Application* - Uses machines, tools, or equipment effectively; uses computers and computer applications to analyze and communicate information in the appropriate format.

33. *Technical Competence* - Uses knowledge that is acquired through formal training or extensive on-the-job experience to perform one's job; works with, understands, and evaluates technical information related to the job; advises others on technical issues.

34. *Memory* - Recalls information that has been presented previously.

35. *Perceptual Speed* - Quickly and accurately sees detail in words, numbers, pictures, and graphs.

36. *Agility* - Bends, stretches, twists, or reaches out with the body, arms, or legs.

37. *Stamina* - Exerts oneself physically over long periods of time without tiring (which may include performing repetitive tasks such as data entry or coding).

38. *Physical Strength* - Exerts maximum muscle force to lift, push, pull, or carry objects; performs moderately laboring work.

39. *Eye-Hand Coordination* - Accurately coordinates one's eyes with one's fingers, wrists, or arms to perform job-related tasks (for example, to move, carry, or manipulate objects).

40. *Spatial Orientation* - Knows one’s location in relation to the environment; determines where other objects are in relation to one’s self (for example, when using a map).

41. *Visual Identification* - Accurately identifies people, animals, or objects based on knowledge of their characteristics.

42. *Peripheral Vision* - Sees objects or movement of objects to one’s side when the eyes are focused forward.

43. *Depth Perception* - Accurately judges which of several objects is closer or farther away from

the observer, or the distance between an object and the observer.

44. *Visual Color Discrimination* - Accurately matches or detects differences between colors, including shades of color and brightness.

Appendix C:

Supplemental Analyses

Concordance of r_{wg} *and* a_{wg}

The analyses were run using both r_{wg} and a_{wg}, but without much attention to their relative performance. That is, do r_{wg} and a_{wg} lead to the same conclusions? Calculated at the level of series-grades, r_{wg} and a_{wg} correlate $r = .79$ for the CT occupations and $r = .65$ for the PA occupations. For the CT occupations, r_{wg} and a_{wg} are both unacceptable (i.e., $r_{wg} < .70$ and $a_{wg} < .80$) in 92.75% of the cases, r_{wg} but not a_{wg} is acceptable in 6.22% of the cases, a_{wg} but not r_{wg} is acceptable in 0.42% of the cases, and both are acceptable in 0.61% of the cases. For the PA occupations, both are unacceptable in 87.45% of the cases, r_{wg} but not a_{wg} is acceptable in 11.60% of the cases, a_{wg} but not r_{wg} is acceptable in 0.30% of the cases, and both are acceptable in 0.40% of the cases. As would be expected, r_{wg} and a_{wg} are strongly correlated, indicating that they tend to rank agreement levels similarly. Also as would be expected, r_{wg}, given the data here, is generally more lenient than a_{wg} (a function of the extreme mean importance ratings). Thus, although the present data preclude a strong test of the concordance between r_{wg} and a_{wg}, these results suggest that they may, in many cases, lead to different conclusions.

Average Interrater Correlations

For the sake of consistency with previous research, average IRCs were calculated for each series-grade. Correlations were calculated between every possible rater pair, the correlations were transformed to *z* scores using Fischer's *r*-to-*z* transformation, averaged, and then translated back to *r* units. Average IRCs were substantially lower for the CT occupations ($M = .25$, $SD = .07$, Min = .11, Max = .41) than for the PA occupations ($M = .56$, $SD = .09$, Min = .30, Max = .73). Note that these results—showing stronger agreement in the PA occupations—

are consistent with the G-analyses, but not with the results of r_{wg} and a_{wg}. For the CT occupations, the average IRC correlated $r = .74$ with multi-item r_{wg} and $r = .39$ with multi-item a_{wg}. For the PA occupations, the correlations were $r = .67$ and $r = .07$, for r_{wg} and a_{wg}, respectively.

Prevalence of Inadmissible $\mathrm{a_{wg}}$ *Values*

The chief limitation of the a_{wg} index is that it cannot be computed for extreme means at small sample sizes. Because the large sample sizes in the present study rendered this limitation largely moot, its practical implications are as yet unknown. Accordingly, treating the entire CT and PA databases as rater populations, I drew 1,000 random samples of size 5 and size 10 for each item, and then computed the percentage of times a_{wg} could not be interpreted. Across all 31 CT and 44 PA occupations, for $n = 5$, $M = .24$, $SD = .21$, $Min = .01$, $Max = .88$, $Med = .19$. For $n = 10$, $M = .04$, $SD = .08$, $Min = 0$, $Max = .50$, $Med = .01$. On average, then, a_{wg} will be uninterpretable 25% of the time when $n = 5$, but only 4% of the time when $n = 10$. As would be expected, the proportion of uninterpretable values is highly correlated with the population mean (calculated from the entire samples) deviation from the scale midpoint, $r = .92$ for $n = 5$ and $r = .65$ for $n = 10$.

Dealing with Disagreement

As the results for Hypothesis 1 show, incumbents often do not agree on the importance of KSAOs. However, practitioners usually do not (and should not) blindly accept the relevance of every rating when setting KSAO requirements. In the absence of any convenient way to check the validity of ratings (which is the norm), outliers are often treated as erroneous or, at least, not representative and are removed from the sample. Table 34 shows agreement at the level of series-grades after removing ratings that fall various distances from the mean. Mean r_{wg} is in the mid .40s and mean a_{wg} in the mid .60s before removing any raters, and both climb to acceptable levels after removing raters (approximately 15%) who score 1.25 – 1.50 standard deviations from the mean. For the CT occupations, more than 95% of r_{wg} and a_{wg} values are acceptable after removing raters who score outside 1.00 standard deviations of the mean. For the PA occupations, removing these raters yields acceptable r_{wg} values in 88% of the cases and acceptable a_{wg} values in 78%. Achieving these levels, though, requires, on average, the removal of approximately one-third of the raters.

If one's sole purpose is to take the average rating as a job's true score, then removing raters to achieve agreement is only a concern to the extent that doing so changes one's conclusion. If ratings are normally distributed, for example, then the mean rating will be the same regardless of how many raters are removed. As skewness increases, however, the greater is the chance that removing raters will change a job's mean score. Table 34 also shows the percentage of items, across all series-grades, whose criticality (defined as a mean of 3.5 or a above; Pollack et al., 2000; Rodriguez et al., 1996) changes after removing raters. This percentage is at or below approximately 10% in all cases, except after removing raters who score .25 standard deviations or more from the mean. The higher percentage in these groups is due to the greater number of groups for which agreement could not be calculated because no raters were left in the sample. Thus, artificially increasing agreement by removing "outliers" has little effect on conclusions of KSAO importance. The question, therefore, is not so much whether removing raters changes conclusions about mean job profiles, but rather the extent to which removing raters obscures meaningful cross-position KSAO requirements.

Table 1

Mean KSAO construct abstractness and item multidimensionality: CT.

Item	Construct Abstractness				Definition Multidimensionality			
	M	*SD*	r_{wg}	a_{wg}	*M*	*SD*	r_{wg}	a_{wg}
1	2.29	0.49	.88	.94	2.86	0.69	.76	.90
2	1.43	0.53	.86	†	2.00	0.82	.67	.81
3	2.57	0.53	.86	.94	3.71	0.76	.71	.86
4	1.43	0.53	.86	†	2.86	0.69	.76	.90
5	2.71	0.76	.71	.88	2.29	0.49	.88	.94
6	4.71	0.49	.88	†	1.86	0.69	.76	.85
7	3.57	0.53	.86	.93	4.14	0.69	.76	.85
8	4.43	0.53	.86	†	2.71	1.25	.21	.66
9	4.57	0.53	.86	†	2.57	0.79	.69	.86
10	3.29	0.76	.71	.88	1.71	0.76	.71	.79
11	1.86	0.69	.76	.85	2.00	1.00	.50	.71
12	1.71	0.76	.71	.79	2.14	0.90	.60	.79
13	1.14	0.38	.93	†	2.00	0.58	.83	.90
14	1.43	0.53	.86	†	1.57	0.79	.69	†
15	2.14	0.90	.60	.79	4.14	0.69	.76	.85
16	2.43	0.79	.69	.86	3.14	0.90	.60	.83
17	3.57	0.53	.86	.93	4.14	0.69	.76	.85
18	2.86	0.90	.60	.83	4.43	0.53	.86	†
19	2.57	0.79	.69	.86	4.14	0.69	.76	.85

20	2.29	0.76	.71	.86	3.71	0.76	.71	.86
21	3.86	0.69	.76	.88	2.00	0.58	.83	.90
22	3.86	0.69	.76	.88	2.14	0.38	.93	.96
23	2.86	0.69	.76	.90	3.57	0.53	.86	.93
24	4.14	0.69	.76	.85	3.43	0.53	.86	.94
25	3.29	0.76	.71	.88	4.29	0.49	.88	.91
26	3.29	0.76	.71	.88	1.43	0.53	.86	†
27	3.71	0.76	.71	.86	3.71	1.11	.38	.70
28	2.57	0.53	.86	.94	4.00	1.00	.50	.71
29	2.71	0.49	.88	.95	3.86	0.90	.60	.79
30	2.71	0.76	.71	.88	2.57	1.27	.19	.64
31	2.43	0.98	.52	.78	2.57	1.27	.19	.64

Note. For items with r_{wg} < .70 and a_{wg} < .80, *M* and *SD* were calculated after removing outliers.

† a_{wg} out-of-range.

Table 2

Mean KSAO construct abstractness and item multidimensionality: PA.

Item	Construct Abstractness				Definition Multidimensionality			
	M	*SD*	r_{wg}	a_{wg}	*M*	*SD*	r_{wg}	a_{wg}
1	2.57	0.79	.69	.86	3.43	0.53	.86	.94
2	1.71	0.76	.71	.79	2.86	0.69	.76	.90
3	1.43	0.53	.86	†	1.86	0.69	.76	.85
4	2.29	0.95	.55	.78	1.71	0.76	.71	.79
5	2.14	1.07	.43	.70	3.71	0.95	.55	.78
6	4.29	0.76	.71	.79	3.29	1.11	.38	.73
7	3.14	0.69	.76	.90	3.57	0.79	.69	.86
8	3.43	0.53	.86	.94	4.14	1.07	.43	.64
9	4.29	0.76	.71	.79	2.86	0.69	.76	.90
10	3.00	0.58	.83	.93	4.00	1.00	.50	.71
11	4.43	0.53	.86	†	2.43	0.79	.69	.86
12	2.71	0.76	.71	.88	3.71	0.76	.71	.86
13	4.29	0.49	.88	.91	3.57	0.53	.86	.93
14	2.57	0.98	.52	.79	3.43	0.98	.52	.79
15	4.14	0.38	.93	.95	2.29	0.76	.71	.86
16	2.86	0.69	.76	.90	4.29	0.76	.71	.79
17	2.86	0.69	.76	.90	3.71	0.95	.55	.78
18	2.57	0.98	.52	.79	4.14	0.69	.76	.85
19	2.14	0.69	.76	.88	1.86	0.69	.76	.85

20	1.86	0.69	.76	.85	4.14	1.07	.43	.64
21	2.43	0.79	.69	.86	4.29	0.49	.88	.91
22	3.86	1.07	.43	.70	2.86	0.69	.76	.90
23	2.71	0.49	.88	.95	2.43	0.53	.86	.93
24	2.71	0.95	.55	.80	2.86	1.07	.43	.75
25	3.71	0.76	.71	.86	3.86	0.69	.76	.88
26	3.14	0.69	.76	.90	4.00	0.82	.67	.81
27	4.43	0.53	.86	†	3.29	0.76	.71	.88
28	2.43	0.98	.52	.78	2.57	1.13	.36	.71
29	2.71	0.95	.55	.80	2.14	0.90	.60	.79
30	3.00	0.58	.83	.93	1.57	0.79	.69	†
31	3.14	0.90	.60	.83	2.57	1.27	.19	.64
32	1.43	0.53	.86	†	3.57	0.98	.52	.78
33	2.00	1.00	.50	.71	4.29	0.76	.71	.79
34	2.00	0.82	.67	.81	1.57	0.79	.69	†
35	1.71	0.76	.71	.79	1.71	0.76	.71	.79
36	1.14	0.38	.93	†	2.00	0.82	.67	.81
37	1.71	0.76	.71	.79	1.43	0.79	.69	.65
38	1.14	0.38	.93	†	1.71	0.76	.71	.79
39	1.29	0.49	.88	†	1.57	0.79	.69	†
40	2.43	0.53	.86	.93	1.43	0.79	.69	†
41	1.71	0.76	.71	.79	1.43	0.79	.69	†
42	1.29	0.49	.88	†	1.29	0.49	.88	†

43	1.29	0.49	.88	†	1.29	0.49	.88	†
44	2.29	0.49	.88	.94	1.29	0.49	.88	†

Note. For items with $r_{wg} < .70$ and $a_{wg} < .80$, *M* and *SD* were calculated after removing outliers.

† a_{wg} out-of-range.

Table 3

Series-grade sample sizes for CT occupations.

	Grade					
Occupational Series	4	5	6	7	8	Total
203—Personnel Clerical and Assistance	146	287	257	405	105	1200
204—Military Personnel Cler. and Tech.	107	110	99	122	68	506
303—Miscellaneous Clerk and Assist.	296	372	403	399	211	1681
305—Mail and File	258	209	141	85	28	721
318—Secretary	310	724	672	526	339	2571
332—Computer Operations	23	92	124	240	141	620
335—Computer Clerk and Assistance	118	175	173	237	126	829
344—Management Cler. and Assistance	31	154	229	373	90	877
503—Financial Clerical and Assistance	86	216	151	210	124	787
525—Accounting Technician	82	226	268	332	125	1033
592—Tax Examining	32	128	165	289	179	793
679—Medical Clerk	119	80	37	34	20	290
986—Legal Clerk and Technician	40	168	230	275	209	922
998—Claims Clerical	43	107	33	24	29	236
1101—General Business and Industry	51	173	154	165	66	609
1105—Purchasing	27	138	138	156	53	512
2005—Supply Clerical and Technician	156	207	173	241	75	852
Total	1925	3566	3447	4113	1988	15039

Table 4

Series-grade sample sizes for PA occupations.

Occupational Series	Grade				
	9	11	12	13	Total
18—Safety and Occupational Health Management	128	217	161	123	629
28—Environmental Protection Specialist	135	192	171	125	623
80—Security Administration	158	185	169	122	634
101—Social Science	109	144	40	34	327
105—Social Insurance Administration	34	169	155	41	399
132—Intelligence	28	69	117	149	363
180—Psychology	69	72	157	184	482
201—Personnel Management	130	262	180	129	701
212—Personnel Staffing	44	166	80	23	313
230—Employee Relations	64	147	116	62	389
235—Employee Development	53	123	114	54	344
301—Miscellaneous Administration and Programs	145	173	204	143	665
334—Computer Specialist	152	295	292	156	895
341—Administrative Officer	149	147	112	84	492
343—Management and Program Analysis	178	194	210	170	752
346—Logistics Management	137	227	224	158	746
391—Telecommunications	124	158	153	89	524
501—Financial Admin. And Programs	116	136	152	95	499
510—Accountant	89	142	148	144	523

511—Auditor	67	147	315	199	728
560—Budget Analysis	146	153	130	111	540
570—Financial Institution Examining	33	50	157	108	348
801—General Engineering	35	181	249	375	840
950—Paralegal Specialist	109	140	87	47	383
996—Veterans Claims Examining	174	123	233	32	562
1001—General Arts and Information	74	84	60	61	279
1035—Public Affairs	100	134	140	96	470
1082—Writing and Editing	109	115	104	51	379
1083—Technical Writing and Editing	99	150	119	23	391
1101—General Business and Industry	142	143	167	133	585
1102—Contract Specialist	149	226	238	159	772
1150—Industrial Specialist	43	194	148	74	459
1165—Loan Specialist	184	170	256	109	719
1170—Realty	135	208	129	98	570
1301—General Physical Science	50	109	150	250	559
1701—General Education and Training	117	82	50	65	314
1801—Civil Aviation Security Specialist	48	110	153	92	403
1810—General Investigator	26	244	166	52	488
1811—Criminal Investigator	60	118	120	148	446
1910—Quality Assurance Specialist	214	286	229	95	824
2003—Supply Program Management	156	184	162	25	527
2010—Inventory Management	155	180	95	27	457

2101—Transportation Specialist	81	88	106	89	364
2130—Traffic Management	77	113	84	23	297
Total	4625	6950	6802	4627	23004

Table 5

Summary of r_{wg} across the 31 CT KSAOs.

	Grade																			
	4				5				6				7				8			
Series	Min	Max	Med	CI%	Min	Max	Med	CI%	Min	Max	Med	CI%	Min	Max	Med	CI%	Min	Max	Med	CI%
203	.01	.71	.43	10	.17	.76	.47	10	.21	.72	.53	13	.16	.77	.57	16	.09	.80	.57	42
204	.28	.68	.47	13	.11	.75	.54	19	.08	.77	.50	23	.26	.77	.55	42	.21	.83	.64	65
303	.11	.62	.37	0	.16	.69	.47	13	.19	.70	.46	6	.18	.76	.50	19	.19	.83	.54	29
305	.19	.60	.35	0	.26	.61	.42	0	.00	.69	.43	13	.02	.72	.53	42	.15	.76	.53	68
318	.26	.68	.47	16	.26	.72	.51	16	.21	.76	.48	19	.28	.77	.52	29	.24	.82	.53	29
332	.01	.65	.42	48	.01	.66	.37	13	.05	.63	.37	3	.21	.66	.44	3	.16	.66	.49	6
335	.17	.64	.44	6	.16	.66	.46	3	.22	.64	.45	13	.22	.66	.47	13	.17	.69	.49	23
344	.17	.74	.42	32	.20	.71	.51	23	.18	.70	.50	13	.19	.70	.57	19	.08	.77	.56	42
503	.32	.66	.47	19	.14	.70	.45	10	.10	.71	.53	19	.15	.69	.51	13	.01	.79	.57	45
525	.13	.66	.38	13	.08	.58	.44	0	.22	.67	.50	16	.17	.68	.46	6	.23	.70	.53	32
592	.00	.81	.32	42	.02	.67	.36	10	.17	.71	.44	13	.21	.76	.46	19	.25	.76	.57	32

679	.13	.78	.45	16	.21	.84	.54	39	.04	.83	.61	71	.23	.80	.67	71	.16	.86	.69	58
986	.15	.71	.42	35	.07	.68	.43	16	.17	.73	.48	19	.17	.78	.49	23	.14	.78	.51	26
998	.10	.78	.55	55	.27	.69	.46	23	.20	.68	.49	52	-.15	.68	.34	45	.06	.80	.42	48
1101	.20	.81	.55	61	.30	.73	.54	32	.34	.75	.55	35	.23	.72	.58	45	.21	.80	.62	58
1105	.15	.79	.58	74	.02	.75	.44	16	.24	.79	.57	29	.24	.83	.52	29	.07	.80	.56	61
2005	.11	.65	.48	3	.19	.62	.48	0	.14	.63	.46	3	.13	.69	.46	16	.33	.75	.60	58

Note. CI% = percentage of values whose 95% confidence interval upper bound is greater than .70.

Table 6

Summary of a_{wg} across the 31 CT KSAOs.

	Grade																			
	4				5				6				7				8			
Series	Min	Max	Med	CI%	Min	Max	Med	CI%	Min	Max	Med	CI%	Min	Max	Med	CI%	Min	Max	Med	CI%
203	.51	.71	.65	0	.59	.74	.68	0	.61	.73	.71	0	.57	.76	.71	0	.54	.77	.70	13
204	.62	.78	.72	26	.55	.77	.71	29	.51	.73	.66	3	.61	.78	.71	13	.60	.77	.71	32
303	.55	.73	.65	0	.58	.77	.69	3	.56	.72	.67	0	.56	.75	.68	0	.57	.75	.68	0
305	.58	.73	.65	0	.62	.77	.69	3	.50	.76	.67	3	.51	.76	.69	23	.52	.84	.71	61
318	.59	.74	.69	0	.62	.73	.68	0	.59	.73	.68	0	.62	.73	.68	0	.58	.74	.68	0
332	.50	.81	.68	58	.50	.74	.65	0	.53	.71	.63	0	.60	.75	.69	0	.58	.80	.70	13
335	.58	.76	.67	13	.58	.76	.70	3	.60	.78	.70	13	.60	.76	.69	3	.59	.80	.70	16
344	.57	.82	.66	35	.57	.80	.71	13	.59	.76	.71	3	.59	.74	.70	0	.54	.78	.70	26
503	.61	.79	.70	16	.57	.75	.68	0	.54	.77	.70	3	.58	.76	.69	6	.51	.75	.69	13
525	.52	.74	.66	3	.54	.76	.68	3	.60	.75	.69	0	.58	.76	.70	0	.56	.79	.71	23
592	.31	.76	.63	35	.50	.76	.66	10	.55	.75	.68	0	.56	.77	.68	3	.59	.75	.69	0

679	.57	.77	.67	10	.61	.77	.70	35	.52	.80	.70	52	.61	.83	.73	84	.34	.89	.71	58
986	.56	.77	.67	35	.54	.76	.66	3	.57	.75	.68	0	.59	.75	.67	0	.57	.73	.67	0
998	.55	.80	.71	61	.63	.77	.69	29	.60	.80	.72	77	.33	.78	.59	29	.47	.79	.63	32
1101	.57	.83	.72	68	.61	.83	.74	26	.66	.78	.73	19	.60	.78	.72	32	.58	.81	.72	42
1105	.59	.88	.74	90	.49	.74	.67	0	.62	.75	.72	6	.60	.78	.68	10	.54	.77	.70	45
2005	.55	.78	.71	19	.59	.77	.70	6	.57	.75	.70	0	.56	.75	.69	0	.64	.81	.73	52

Note. CI% = percentage of values whose 95% confidence interval upper bound is greater than .80.

Table 7

Summary of r_{wg} *across the 44 PA KSAOs.*

	Grade															
	9				11				12				13			
Series	Min	Max	Med	CI%	Min	Max	Med	CI%	Min	Max	Med	CI%	Min	Max	Med	CI%
18	.03	.75	.45	23	.05	.85	.44	16	.17	.84	.50	25	.07	.86	.50	30
28	-.06	.82	.40	20	.03	.79	.47	16	.09	.85	.50	25	.11	.92	.53	39
80	-.03	.68	.36	9	-.05	.79	.36	14	.03	.86	.42	18	.11	.82	.44	27
101	-.05	.87	.42	32	-.13	.81	.36	32	-.24	.95	.45	52	-.21	.90	.48	48
105	-.33	.97	.38	45	-.12	.93	.54	45	-.14	.87	.42	27	.01	.99	.54	59
132	-.31	.80	.37	43	-.07	.80	.46	45	-.18	.79	.39	18	-.15	.80	.49	25
180	.00	.91	.39	41	-.04	.87	.42	36	-.12	.88	.46	32	-.04	.83	.48	30
201	-.31	.82	.46	23	-.21	.87	.54	27	-.21	.88	.53	39	-.16	.93	.54	50
212	-.31	.92	.56	55	-.27	.85	.56	41	-.27	.90	.58	57	-.17	.98	.55	68
230	-.47	.87	.46	43	-.21	.89	.51	41	-.12	.90	.58	57	-.08	.94	.42	48
235	-.14	.78	.44	39	-.03	.86	.52	34	-.22	.82	.49	39	.02	.89	.61	66

301	-.08	.75	.37	14	-.19	.70	.40	14	-.16	.81	.42	23	.02	.77	.49	36
334	.14	.66	.40	16	.09	.75	.47	16	-.03	.70	.47	14	.07	.80	.59	36
341	.01	.75	.45	18	.00	.77	.48	20	.16	.84	.56	48	.21	.89	.63	61
343	-.18	.71	.46	16	-.09	.71	.48	27	-.02	.76	.49	18	-.13	.88	.57	48
346	.11	.75	.47	18	.01	.70	.46	14	.03	.78	.50	23	.20	.79	.55	39
391	-.05	.72	.43	16	-.10	.68	.47	25	.00	.76	.46	23	-.08	.81	.46	36
501	-.26	.75	.54	34	-.28	.74	.45	20	-.14	.77	.52	23	-.18	.88	.54	48
510	-.07	.73	.45	16	-.22	.73	.55	32	-.13	.83	.48	23	-.21	.83	.57	36
511	-.22	.86	.46	36	.00	.86	.51	30	.09	.86	.53	34	.00	.87	.54	41
560	.12	.76	.53	23	-.18	.73	.46	20	-.06	.77	.51	39	.16	.87	.56	50
570	-.54	.91	.66	89	-.27	1.00	.62	64	-.40	.90	.64	55	-.29	.86	.62	59
801	-.16	.70	.45	48	.01	.86	.52	20	.15	.80	.48	23	.22	.85	.55	25
950	-.03	.79	.38	20	.00	.80	.42	14	-.22	.97	.35	34	-.21	1.00	.47	55
996	-.08	.93	.31	23	-.13	.95	.37	30	-.24	.94	.35	32	-.38	.96	.45	52
1001	-.20	.68	.32	14	-.32	.78	.32	14	-.50	.75	.15	14	-.05	.73	.37	18
1035	-.01	.85	.45	30	-.10	.90	.46	32	.07	.88	.50	36	-.01	.98	.59	48

1082	.06	.89	.42	27	-.14	.88	.37	23	-.14	.85	.28	18	-.26	.87	.32	41
1083	-.09	.78	.39	25	.04	.85	.39	18	.04	.90	.48	20	-.30	.98	.46	61
1101	-.07	.69	.36	5	.05	.69	.45	11	-.12	.75	.46	16	.01	.83	.56	39
1102	-.06	.81	.43	14	.01	.84	.53	20	-.04	.81	.50	27	.13	.90	.58	43
1150	-.10	.78	.39	39	.03	.79	.46	18	.09	.73	.44	27	.20	.90	.55	43
1165	-.21	.79	.39	18	-.13	.79	.48	25	-.11	.83	.53	30	-.12	.91	.51	34
1170	.00	.81	.49	30	-.01	.80	.47	20	.04	.83	.52	39	.01	.87	.52	36
1301	-.09	.80	.47	39	-.01	.79	.47	16	.12	.83	.47	27	.12	.89	.52	23
1701	-.20	.83	.33	25	-.22	.87	.37	36	-.07	.82	.43	41	.09	.95	.54	48
1801	-.32	.77	.38	32	-.08	.81	.29	20	.07	.77	.43	18	-.17	.92	.38	27
1810	-.05	.95	.36	55	-.19	.88	.34	20	-.14	.86	.46	27	-.01	.91	.45	36
1811	-.10	.96	.41	39	.06	.95	.43	32	.01	.81	.40	39	.04	.83	.41	25
1910	.03	.83	.45	18	-.02	.82	.43	16	.13	.80	.51	30	.12	.82	.48	36
2003	-.05	.78	.49	14	.00	.74	.44	16	.02	.77	.45	16	-.20	.81	.52	77
2010	-.06	.70	.47	18	-.10	.67	.45	9	.06	.77	.50	32	-.19	.92	.56	77
2101	-.08	.70	.36	11	-.07	.75	.44	27	-.08	.76	.37	16	.03	.82	.35	25

2130	-.20	.79	.44	23	-.10	.78	.44	23	.04	.77	.46	25	.14	.88	.57	75

Note. CI% = percentage of values whose 95% confidence interval upper bound is greater than .70.

Table 8

Summary of a_{wg} across the 44 PA KSAOs.

	Grade															
	9				11				12				13			
Series	Min	Max	Med	CI%	Min	Max	Med	CI%	Min	Max	Med	CI%	Min	Max	Med	CI%
18	.50	.79	.67	11	.52	.76	.68	5	.57	.77	.71	18	.54	.80	.69	20
28	.43	.76	.65	5	.50	.77	.67	11	.52	.75	.68	0	.50	.78	.68	16
80	.46	.72	.60	0	.37	.74	.62	0	.47	.77	.66	2	.46	.78	.66	14
101	.42	.75	.60	5	.39	.74	.60	0	.32	.78	.56	34	.41	.78	.62	48
105	.22	.78	.60	36	.25	.75	.60	2	.39	.71	.60	2	.38	.90	.65	45
132	.36	.79	.61	41	.43	.77	.62	11	.38	.73	.59	2	.37	.75	.64	2
180	.37	.77	.57	14	.40	.76	.62	5	.39	.75	.63	0	.42	.77	.60	2
201	.35	.73	.63	0	.39	.73	.64	0	.40	.77	.66	7	.40	.76	.64	14
212	.26	.80	.67	55	.34	.78	.66	5	.37	.78	.63	18	.38	.88	.69	73
230	.23	.77	.64	23	.36	.77	.67	9	.37	.77	.66	11	.39	.75	.58	23
235	.42	.77	.67	30	.46	.80	.69	34	.40	.78	.65	9	.47	.82	.71	68

301	.42	.74	.64	2	.40	.73	.63	0	.41	.75	.63	2	.43	.78	.66	11
334	.47	.76	.65	2	.45	.75	.65	0	.41	.73	.63	0	.49	.78	.68	9
341	.47	.75	.66	0	.49	.78	.65	2	.47	.80	.70	23	.48	.81	.74	64
343	.39	.76	.63	2	.44	.76	.67	7	.38	.76	.67	5	.44	.76	.68	7
346	.53	.77	.68	11	.50	.74	.67	0	.51	.77	.67	2	.52	.77	.71	20
391	.45	.76	.67	7	.45	.77	.66	11	.46	.77	.67	5	.41	.77	.65	18
501	.37	.78	.69	20	.35	.76	.63	11	.39	.77	.67	2	.35	.79	.64	23
510	.43	.76	.65	9	.39	.77	.69	11	.44	.74	.67	5	.40	.77	.68	14
511	.30	.76	.61	20	.48	.76	.67	7	.50	.75	.67	0	.48	.75	.67	2
560	.50	.80	.69	18	.41	.72	.64	0	.45	.77	.66	7	.43	.81	.70	27
570	.25	.84	.76	82	.34	1.00	.72	70	.29	.79	.71	39	.34	.85	.74	59
801	.30	.83	.66	45	.49	.80	.68	7	.48	.77	.70	5	.51	.79	.71	5
950	.33	.74	.62	2	.39	.75	.64	2	.35	.76	.53	11	.27	1.00	.63	30
996	.27	.73	.54	0	.34	.73	.55	0	.27	.71	.55	2	.16	.81	.61	43
1001	.41	.74	.59	2	.35	.76	.60	7	.24	.77	.54	7	.35	.76	.60	7
1035	.47	.75	.64	2	.44	.76	.64	2	.46	.75	.62	2	.46	.76	.67	14

1082	.41	.79	.64	9	.36	.72	.58	2	.30	.70	.58	2	.33	.73	.55	23
1083	.41	.77	.62	9	.44	.74	.60	2	.49	.75	.67	2	.36	.83	.61	43
1101	.40	.75	.62	0	.45	.77	.69	11	.41	.76	.66	2	.49	.81	.67	14
1102	.42	.73	.64	0	.43	.78	.67	2	.43	.74	.64	0	.41	.78	.68	9
1150	.33	.76	.63	30	.44	.78	.67	7	.47	.77	.68	20	.57	.85	.70	41
1165	.35	.70	.62	0	.40	.74	.67	0	.43	.76	.68	0	.40	.78	.68	16
1170	.46	.76	.69	7	.46	.78	.67	2	.48	.79	.69	18	.45	.77	.69	25
1301	.46	.81	.68	32	.47	.79	.68	18	.51	.79	.69	9	.48	.76	.67	2
1701	.38	.71	.56	2	.37	.76	.62	11	.38	.79	.64	25	.40	.80	.68	32
1801	.35	.79	.62	16	.43	.77	.61	7	.44	.74	.63	0	.32	.80	.60	2
1810	.39	.85	.61	43	.40	.73	.59	0	.35	.73	.59	0	.43	.79	.64	25
1811	.40	.81	.64	20	.50	.78	.66	14	.50	.75	.65	7	.51	.78	.68	5
1910	.49	.77	.66	2	.42	.75	.69	0	.54	.79	.70	7	.47	.79	.69	23
2003	.41	.78	.68	2	.41	.76	.66	2	.41	.76	.65	7	.29	.80	.68	73
2010	.39	.79	.67	7	.41	.75	.62	2	.41	.77	.65	11	.42	.85	.73	86
2101	.45	.76	.64	11	.47	.81	.68	20	.41	.74	.63	2	.46	.82	.62	5

2130	.39	.78	.66	16	.44	.74	.66	2	.40	.77	.64	16	.45	.83	.71	82

Note. CI% = percentage of values whose 95% confidence interval upper bound is greater than .80.

Table 9

Multi-item agreement on CT KSAOs.

	r_{wg}										a_{wg}									
	Grade										Grade									
	4		5		6		7		8		4		5		6		7		8	
Series	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
203	.40	.18	.46	.15	.50	.14	.52	.15	.52	.20	.64	.05	.67	.04	.69	.03	.71	.04	.69	.05
204	.47	.11	.50	.14	.47	.16	.55	.15	.56	.20	.70	.04	.70	.05	.66	.04	.70	.04	.69	.05
303	.39	.13	.46	.13	.44	.14	.49	.16	.51	.18	.65	.04	.69	.04	.66	.04	.67	.05	.68	.05
305	.37	.11	.44	.10	.43	.16	.50	.16	.50	.16	.65	.04	.69	.03	.66	.06	.68	.06	.70	.07
318	.48	.13	.49	.15	.49	.16	.51	.16	.52	.18	.69	.04	.68	.03	.68	.04	.68	.03	.68	.04
332	.39	.17	.38	.17	.36	.14	.44	.12	.46	.14	.68	.08	.64	.05	.63	.04	.69	.04	.70	.05
335	.42	.13	.44	.12	.47	.12	.47	.14	.47	.15	.68	.05	.68	.05	.70	.05	.68	.04	.69	.05
344	.43	.16	.49	.13	.50	.14	.51	.15	.49	.20	.66	.06	.70	.05	.70	.04	.69	.04	.68	.07
503	.49	.10	.44	.14	.49	.15	.48	.15	.50	.20	.70	.04	.67	.05	.69	.05	.68	.05	.67	.06
525	.40	.14	.42	.13	.49	.13	.46	.14	.51	.13	.65	.05	.67	.05	.69	.04	.68	.05	.71	.05

592	.33	.19	.38	.15	.45	.15	.48	.16	.53	.15	.61	.08	.65	.06	.67	.05	.68	.05	.68	.04
679	.45	.16	.54	.15	.56	.21	.58	.16	.61	.20	.67	.05	.70	.05	.70	.06	.73	.06	.70	.09
986	.43	.17	.42	.17	.46	.17	.48	.17	.46	.19	.66	.05	.65	.05	.67	.05	.67	.04	.66	.05
998	.50	.18	.48	.12	.49	.14	.31	.23	.42	.17	.69	.06	.70	.04	.71	.05	.60	.10	.63	.08
1101	.54	.18	.55	.12	.55	.12	.55	.13	.56	.20	.72	.07	.73	.04	.72	.03	.72	.05	.71	.06
1105	.55	.15	.45	.16	.54	.14	.52	.16	.53	.17	.75	.06	.66	.05	.71	.04	.69	.04	.70	.05
2005	.46	.13	.46	.11	.45	.12	.48	.13	.58	.12	.70	.05	.70	.04	.69	.04	.69	.04	.72	.04

Table 10

Multi-item agreement on CT KSAOs.

	r_{wg}								a_{wg}							
	Grade								Grade							
	9		11		12		13		9		11		12		13	
Series	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
18	.45	.19	.45	.20	.50	.16	.51	.16	.67	.07	.66	.06	.69	.05	.69	.05
28	.38	.24	.45	.17	.49	.17	.52	.20	.63	.09	.67	.07	.66	.06	.65	.08
80	.37	.17	.36	.20	.44	.21	.45	.20	.60	.07	.60	.09	.63	.08	.64	.09
101	.42	.24	.40	.24	.38	.28	.45	.26	.59	.09	.59	.09	.57	.10	.61	.09
105	.38	.31	.48	.23	.42	.23	.52	.24	.56	.13	.57	.11	.59	.08	.64	.09
132	.33	.28	.41	.26	.35	.24	.43	.25	.59	.12	.61	.09	.58	.10	.62	.10
180	.39	.27	.44	.23	.46	.23	.46	.23	.57	.11	.60	.09	.61	.11	.59	.08
201	.42	.23	.48	.21	.52	.23	.53	.26	.60	.10	.62	.09	.64	.09	.62	.09
212	.51	.26	.50	.23	.52	.27	.53	.26	.65	.12	.63	.10	.63	.08	.66	.11
230	.45	.26	.49	.23	.57	.23	.46	.27	.62	.10	.62	.11	.63	.10	.58	.10

235	.43	.23	.51	.20	.49	.21	.56	.20	.65	.09	.67	.10	.64	.09	.68	.09
301	.37	.21	.37	.23	.42	.24	.46	.21	.60	.09	.60	.10	.61	.10	.64	.10
334	.41	.16	.43	.17	.41	.19	.52	.17	.64	.07	.63	.08	.60	.09	.66	.08
341	.43	.18	.46	.19	.55	.17	.61	.17	.64	.08	.64	.08	.68	.08	.71	.08
343	.41	.21	.45	.19	.45	.20	.54	.24	.62	.08	.64	.09	.62	.11	.66	.08
346	.44	.17	.42	.18	.49	.17	.55	.14	.66	.07	.64	.08	.65	.07	.68	.07
391	.41	.19	.41	.23	.42	.20	.42	.23	.65	.07	.65	.09	.64	.08	.63	.10
501	.46	.20	.40	.23	.46	.20	.51	.24	.66	.09	.61	.11	.63	.10	.64	.09
510	.42	.17	.49	.21	.46	.18	.50	.22	.63	.09	.66	.09	.64	.08	.65	.08
511	.42	.26	.49	.21	.52	.19	.54	.19	.60	.10	.64	.07	.65	.07	.65	.06
560	.49	.16	.43	.21	.49	.19	.55	.18	.67	.08	.61	.09	.63	.08	.67	.10
570	.60	.25	.60	.27	.59	.23	.59	.21	.68	.17	.72	.11	.69	.10	.70	.12
801	.40	.21	.47	.18	.49	.16	.53	.15	.64	.11	.66	.07	.66	.08	.68	.07
950	.39	.21	.40	.19	.39	.30	.50	.30	.58	.12	.60	.10	.54	.11	.60	.14
996	.34	.25	.37	.27	.41	.29	.43	.28	.52	.12	.53	.09	.53	.10	.57	.16
1001	.28	.21	.34	.25	.18	.27	.34	.18	.58	.08	.60	.10	.52	.11	.59	.09

1035	.45	.21	.48	.23	.49	.21	.54	.21	.63	.06	.63	.07	.62	.07	.64	.08
1082	.44	.21	.35	.25	.33	.22	.36	.28	.62	.09	.57	.09	.56	.09	.54	.09
1083	.39	.21	.40	.19	.47	.17	.44	.30	.60	.10	.59	.08	.64	.07	.58	.09
1101	.33	.20	.41	.18	.43	.20	.52	.19	.61	.08	.65	.08	.62	.10	.66	.08
1102	.41	.21	.49	.20	.47	.20	.55	.18	.61	.09	.65	.09	.62	.09	.65	.10
1150	.37	.22	.42	.20	.47	.18	.55	.18	.61	.10	.64	.09	.66	.09	.69	.07
1165	.37	.24	.46	.21	.49	.19	.49	.21	.59	.09	.64	.08	.66	.08	.64	.11
1170	.47	.20	.46	.19	.49	.22	.49	.18	.67	.07	.66	.06	.66	.09	.66	.09
1301	.41	.24	.44	.19	.48	.17	.50	.17	.65	.10	.67	.08	.67	.07	.66	.07
1701	.36	.28	.40	.29	.43	.22	.52	.23	.57	.08	.60	.11	.62	.10	.65	.11
1801	.34	.29	.32	.26	.39	.21	.39	.27	.60	.10	.59	.09	.62	.08	.58	.10
1810	.40	.28	.39	.25	.41	.26	.44	.24	.61	.12	.58	.08	.58	.09	.64	.10
1811	.42	.28	.46	.21	.44	.21	.43	.21	.63	.09	.67	.07	.65	.06	.66	.06
1910	.42	.21	.42	.22	.49	.18	.49	.19	.65	.08	.65	.09	.68	.08	.66	.08
2003	.43	.20	.43	.20	.44	.19	.46	.25	.64	.09	.62	.11	.62	.10	.63	.14
2010	.41	.21	.39	.19	.45	.17	.52	.22	.63	.10	.60	.10	.63	.09	.68	.12

2101	.36	.17	.40	.23	.36	.22	.37	.19	.63	.07	.66	.08	.62	.08	.61	.08
2130	.38	.24	.43	.20	.44	.18	.55	.18	.62	.11	.64	.08	.63	.10	.69	.08

Table 11

Mean agreement at different levels of aggregation.

	r_{wg}			a_{wg}		
Aggregation	*M*	*SD*	*N*	*M*	*SD*	*N*
CT						
S	.47	.15	527	.68	.04	527
SG	.48	.19	2573	.69	.07	2537
SGA	.49	.30	10787	.72	.13	9791
SGAL	.49	.33	19654	.72	.14	17294
PA						
S	.44	.21	1936	.63	.09	1936
SG	.46	.23	7744	.64	.10	7666
SGA	.47	.33	42592	.67	.15	37659

Note. S = series; SG = series-grade; SGA = series-grade-agency; SGAL = series-grade-agency-location. *N* is lower for a_{wg} due to uninterpretable values.

Table 12

Summary of r_{wg} using estimated population variances: CT.

					r_{wg}							
	Reference Variance				Obs		Med		25%		75%	
Item	Obs	Med	25%	75%	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	0.95	0.93	0.75	1.15	0.03	0.28	0.01	0.28	-0.24	0.35	0.19	0.23
2	1.05	1.04	0.80	1.25	0.05	0.30	0.04	0.30	-0.25	0.39	0.20	0.25
3	1.02	1.00	0.80	1.21	0.03	0.19	0.01	0.19	-0.24	0.24	0.18	0.16
4	0.95	0.91	0.75	1.10	0.02	0.21	-0.02	0.22	-0.24	0.26	0.16	0.18
5	1.54	1.52	1.29	1.79	0.15	0.19	0.14	0.19	-0.01	0.22	0.27	0.16
6	1.06	1.04	0.85	1.27	0.06	0.18	0.05	0.18	-0.17	0.22	0.22	0.15
7	1.09	1.08	0.87	1.31	0.06	0.21	0.05	0.21	-0.18	0.27	0.22	0.18
8	1.23	1.21	1.00	1.43	0.02	0.15	0.00	0.15	-0.21	0.18	0.16	0.12
9	1.41	1.40	1.16	1.63	0.02	0.12	0.00	0.12	-0.20	0.15	0.15	0.11
10	0.85	0.83	0.66	1.00	0.01	0.16	-0.01	0.17	-0.27	0.21	0.16	0.14
11	1.48	1.46	1.21	1.71	0.02	0.13	0.01	0.13	-0.20	0.16	0.15	0.11
12	1.25	1.21	1.00	1.46	-0.01	0.14	-0.04	0.15	-0.26	0.18	0.14	0.12
13	1.52	1.52	1.25	1.82	0.08	0.18	0.08	0.18	-0.12	0.22	0.23	0.15
14	1.47	1.46	1.21	1.71	0.03	0.12	0.02	0.12	-0.18	0.15	0.17	0.11
15	1.23	1.21	0.99	1.42	0.02	0.15	0.00	0.16	-0.22	0.19	0.15	0.13
16	0.81	0.77	0.62	0.98	0.05	0.20	0.00	0.21	-0.24	0.26	0.22	0.16
17	1.10	1.08	0.88	1.29	0.02	0.18	0.00	0.18	-0.22	0.22	0.17	0.15
18	1.01	0.99	0.79	1.20	0.04	0.17	0.02	0.17	-0.23	0.21	0.19	0.14

19	1.44	1.42	1.19	1.68	0.02	0.13	0.00	0.13	-0.20	0.16	0.16	0.11
20	1.55	1.52	1.29	1.82	0.06	0.17	0.04	0.18	-0.13	0.21	0.20	0.15
21	0.69	0.66	0.54	0.83	0.00	0.19	-0.04	0.20	-0.28	0.25	0.17	0.16
22	0.64	0.58	0.47	0.78	0.01	0.25	-0.09	0.28	-0.34	0.34	0.19	0.21
23	0.69	0.66	0.54	0.83	0.01	0.26	-0.04	0.27	-0.28	0.34	0.17	0.22
24	0.80	0.77	0.62	0.94	0.00	0.22	-0.05	0.23	-0.29	0.29	0.15	0.19
25	0.75	0.73	0.59	0.89	0.01	0.15	-0.02	0.15	-0.26	0.19	0.17	0.12
26	0.78	0.75	0.62	0.91	0.02	0.15	-0.02	0.16	-0.23	0.19	0.16	0.13
27	1.25	1.22	1.00	1.48	0.04	0.19	0.01	0.20	-0.20	0.24	0.19	0.16
28	1.16	1.14	0.94	1.40	0.06	0.18	0.04	0.18	-0.17	0.22	0.22	0.15
29	0.97	0.94	0.77	1.14	0.05	0.18	0.02	0.18	-0.20	0.22	0.19	0.15
30	1.19	1.19	0.96	1.40	0.03	0.17	0.03	0.17	-0.20	0.21	0.18	0.14
31	0.78	0.75	0.57	0.98	0.07	0.27	0.02	0.28	-0.28	0.37	0.25	0.22

Note. Obs = observed variance calculated on the entire dataset. Med = median of Monte Carlo estimated variances. 25% = the 25th percentile of Monte Carlo estimated variances. 75% = the 75th percentile of Monte Carlo estimated variances.

Table 13

Summary of r_{wg} using estimated population variances: PA.

					r_{wg}							
	Reference Variance				Obs		Med		25%		75%	
Item	Obs	Med	25%	75%	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	0.45	0.38	0.27	0.57	0.04	0.41	-0.13	0.49	-0.57	0.67	0.25	0.32
2	0.60	0.57	0.43	0.75	0.11	0.40	0.06	0.42	-0.24	0.55	0.28	0.32
3	1.60	1.59	1.31	1.84	0.19	0.20	0.18	0.20	0.01	0.24	0.29	0.17
4	1.99	1.99	1.71	2.26	0.20	0.16	0.20	0.16	0.07	0.19	0.29	0.14
5	0.84	0.79	0.58	1.06	0.09	0.39	0.04	0.41	-0.31	0.56	0.29	0.30
6	1.28	1.25	1.04	1.50	0.09	0.17	0.07	0.18	-0.11	0.21	0.23	0.15
7	1.06	1.01	0.80	1.25	0.01	0.25	-0.04	0.26	-0.31	0.33	0.16	0.21
8	0.92	0.87	0.67	1.12	0.04	0.28	-0.01	0.29	-0.30	0.38	0.21	0.23
9	0.91	0.87	0.67	1.10	0.03	0.28	-0.01	0.29	-0.30	0.38	0.21	0.23
10	0.89	0.83	0.66	1.10	0.02	0.32	-0.05	0.34	-0.33	0.42	0.21	0.25
11	1.74	1.73	1.46	2.03	0.08	0.15	0.08	0.15	-0.09	0.18	0.21	0.13
12	1.04	1.00	0.83	1.21	0.01	0.18	-0.03	0.19	-0.24	0.22	0.15	0.15
13	0.92	0.89	0.69	1.12	0.03	0.21	0.00	0.22	-0.29	0.28	0.20	0.17
14	0.91	0.87	0.68	1.10	0.04	0.25	-0.01	0.26	-0.29	0.33	0.20	0.21
15	0.62	0.58	0.45	0.78	0.03	0.34	-0.04	0.36	-0.33	0.47	0.23	0.27
16	0.61	0.57	0.46	0.73	0.02	0.23	-0.05	0.24	-0.29	0.30	0.18	0.19
17	0.69	0.66	0.47	0.83	0.07	0.31	0.03	0.32	-0.36	0.44	0.23	0.25
18	0.89	0.84	0.66	1.08	0.01	0.25	-0.05	0.27	-0.34	0.34	0.18	0.21

19	0.69	0.66	0.54	0.83	0.03	0.19	-0.01	0.20	-0.24	0.25	0.20	0.16
20	2.05	2.05	1.73	2.37	0.19	0.23	0.19	0.23	0.05	0.27	0.30	0.20
21	2.05	2.05	1.75	2.34	0.06	0.16	0.06	0.16	-0.10	0.19	0.18	0.14
22	1.75	1.73	1.46	2.01	0.07	0.17	0.06	0.18	-0.12	0.21	0.19	0.15
23	1.49	1.48	1.21	1.73	0.06	0.17	0.05	0.18	-0.16	0.22	0.19	0.15
24	1.51	1.48	1.16	1.84	0.09	0.33	0.07	0.33	-0.19	0.43	0.25	0.27
25	1.11	1.08	0.89	1.31	0.04	0.18	0.01	0.18	-0.19	0.22	0.18	0.15
26	1.46	1.43	1.21	1.71	0.06	0.14	0.04	0.15	-0.14	0.17	0.20	0.12
27	1.33	1.31	1.08	1.54	0.04	0.16	0.03	0.16	-0.18	0.20	0.17	0.14
28	1.55	1.53	1.29	1.82	0.08	0.18	0.07	0.18	-0.10	0.22	0.22	0.15
29	1.85	1.84	1.57	2.11	0.07	0.14	0.06	0.15	-0.10	0.17	0.18	0.13
30	1.03	1.00	0.79	1.25	0.03	0.20	-0.01	0.20	-0.27	0.26	0.20	0.16
31	0.85	0.83	0.68	1.00	0.01	0.16	-0.01	0.16	-0.23	0.19	0.16	0.13
32	1.11	1.08	0.83	1.32	0.03	0.21	0.01	0.22	-0.29	0.28	0.18	0.18
33	1.09	1.05	0.80	1.31	0.02	0.28	-0.02	0.29	-0.34	0.38	0.19	0.23
34	0.87	0.84	0.68	1.01	0.01	0.15	-0.02	0.15	-0.26	0.19	0.15	0.13
35	1.30	1.29	1.06	1.52	0.00	0.18	0.00	0.19	-0.22	0.23	0.15	0.16
36	1.32	1.29	0.98	1.63	0.13	0.27	0.10	0.27	-0.18	0.36	0.29	0.22
37	1.46	1.43	1.15	1.78	0.08	0.21	0.06	0.21	-0.18	0.26	0.24	0.17
38	0.88	0.80	0.51	1.19	0.14	0.43	0.06	0.47	-0.46	0.73	0.37	0.32
39	1.47	1.46	1.14	1.80	0.08	0.24	0.08	0.24	-0.18	0.31	0.25	0.20
40	1.45	1.43	1.10	1.78	0.16	0.30	0.15	0.30	-0.11	0.39	0.31	0.24
41	1.58	1.57	1.22	1.92	0.10	0.27	0.10	0.27	-0.16	0.35	0.27	0.22

42	1.26	1.22	0.87	1.62	0.13	0.32	0.10	0.34	-0.26	0.47	0.32	0.25
43	1.22	1.19	0.85	1.57	0.15	0.35	0.12	0.36	-0.22	0.50	0.34	0.27
44	1.36	1.36	1.00	1.73	0.16	0.39	0.15	0.39	-0.15	0.53	0.34	0.30

Note. Obs = observed variance calculated on the entire dataset. Med = median of Monte Carlo estimated variances. 25% = the 25th percentile of Monte Carlo estimated variances. 75% = the 75th percentile of Monte Carlo estimated variances.

Table 14

Generalizability analysis for item × rater design: CT.

		Variance Components			Φ		*n* for
Series	*n*	i	r	ir,e	*n* = 5	*n* = 20	Φ = .80
203	390	.27	.37	.59	.58	.85	15
204	209	.24	.33	.55	.58	.85	15
303	621	.18	.33	.61	.49	.79	21
305	290	.10	.41	.60	.34	.67	39
318	925	.27	.31	.57	.61	.86	14
332	265	.13	.44	.65	.38	.71	34
335	363	.13	.34	.63	.40	.73	30
344	350	.22	.31	.56	.56	.84	16
503	283	.23	.30	.63	.55	.83	17
525	361	.19	.37	.58	.50	.80	21
592	284	.24	.34	.59	.57	.84	16
679	109	.18	.30	.57	.51	.81	20
986	292	.30	.32	.65	.60	.86	14
998	80	.22	.29	.62	.55	.83	17
1101	288	.26	.30	.56	.61	.86	14
1105	192	.26	.33	.52	.61	.86	13
2005	354	.11	.39	.53	.36	.70	35

Note. *n* = number of raters. i = item. r = rater. ir,e = item × rater and error. Φ for *N* = *n* is calculated using the observed *n*. Φ for *n* = 5 is calculated assuming 5 raters. The *n* needed to obtain $\Phi = .80$ was calculated by rounding up the estimate.

Table 15

Generalizability analysis for item × rater design: PA.

		Variance Components			Φ		*n* for
Series	*n*	i	r	ir,e	*n* = 5	*n* = 20	Φ = .80
18	570	.53	.36	.70	.71	.91	9
28	579	.72	.32	.80	.76	.93	7
80	571	.86	.39	.80	.78	.93	6
101	297	.98	.38	.84	.80	.94	5
105	358	1.33	.28	.87	.85	.96	4
132	329	.86	.34	.88	.78	.93	6
180	443	1.14	.30	.81	.84	.95	4
201	641	1.26	.27	.77	.86	.96	4
212	288	1.35	.23	.73	.87	.97	3
230	356	1.37	.23	.78	.87	.96	3
235	314	1.00	.26	.72	.83	.95	4
301	591	.86	.33	.86	.78	.94	6
334	817	.92	.31	.82	.80	.94	5
341	448	.92	.29	.72	.82	.95	5
343	682	.97	.29	.79	.82	.95	5
346	692	.83	.33	.73	.80	.94	6
391	469	.57	.34	.83	.71	.91	9
501	454	.97	.29	.80	.82	.95	5
510	482	.97	.30	.76	.82	.95	5

511	680	1.15	.28	.72	.85	.96	4
560	486	.98	.30	.73	.83	.95	5
570	333	1.27	.21	.61	.89	.97	3
801	784	.88	.28	.72	.81	.95	5
950	347	1.15	.31	.91	.82	.95	5
996	518	1.32	.32	.96	.84	.95	4
1001	256	.61	.42	1.02	.68	.89	10
1035	419	1.08	.27	.74	.84	.96	4
1082	350	1.03	.29	.95	.81	.94	5
1083	361	1.04	.33	.82	.82	.95	5
1101	539	.77	.36	.82	.77	.93	7
1102	710	1.10	.29	.74	.84	.96	4
1150	422	.79	.36	.74	.78	.93	6
1165	662	.93	.33	.76	.81	.94	5
1170	512	.75	.32	.75	.78	.93	6
1301	531	.78	.29	.78	.79	.94	6
1701	284	.88	.33	.84	.79	.94	6
1801	369	.74	.36	.95	.74	.92	8
1810	450	1.11	.29	.94	.82	.95	5
1811	408	.38	.35	.77	.63	.87	12
1910	756	.55	.37	.76	.71	.91	9
2003	482	.90	.32	.83	.80	.94	6
2010	413	.88	.33	.83	.79	.94	6

2101	332	.44	.46	.84	.63	.87	12
2130	259	.78	.36	.79	.77	.93	6

Note. *n* = number of raters. i = item. r = rater. ir,e = item × rater and error. Φs are calculated assuming either 5 or 20 raters. The *n* needed to obtain Φ = .80 was calculated by rounding up the estimate.

Table 16

Generalizability analysis for item × rater(grade) design: CT.

		Variance Components					Φ		*n* for
Series	*N*	i	g	ig	r:g	irg,e	*n* = 5	*n* = 20	Φ = .80
203	390	.25	.01	.00	.36	.59	.56	.81	19
204	209	.23	.04	.02	.30	.54	.50	.68	-104
303	621	.18	.01	.01	.33	.60	.47	.73	36
305	290	.12	.02	.01	.40	.59	.34	.61	536
318	925	.26	.00	.00	.31	.56	.59	.84	15
332	265	.11	.02	.01	.42	.65	.31	.57	818
335	363	.13	-.00	.01	.35	.63	.39	.70	40
344	350	.20	.00	.00	.31	.56	.52	.80	20
503	283	.22	-.00	.01	.30	.62	.53	.80	21
525	361	.17	-.00	.01	.37	.57	.47	.76	27
592	284	.23	.00	.01	.33	.58	.53	.78	24
679	109	.19	.00	.02	.29	.55	.51	.76	30
986	292	.29	-.01	.00	.32	.65	.60	.85	14
998	80	.25	.01	.01	.27	.61	.56	.79	22
1101	288	.27	.00	.02	.30	.55	.59	.82	17
1105	192	.25	-.02	.01	.34	.52	.58	.83	16
2005	354	.11	.01	.01	.38	.52	.34	.62	156

Note. *n* = number of raters. i = item. g = grade. r = rater. irg,e = item × rater(grade) and error. Φs are calculated assuming either 5 or 20 raters. The *n* needed to obtain Φ = .80 was calculated by rounding up the estimate. All estimates assume a single grade and were calculated after setting negative variance components to 0.

Table 17

Generalizability analysis for item × rater(grade) design: PA.

		Variance Components					Φ		*n* for
Series	*N*	i	g	ig	r:g	irg,e	*n* = 5	*n* = 20	Φ = .80
18	570	.53	.00	.01	.36	.69	.70	.89	9
28	579	.72	.01	.04	.31	.77	.73	.87	9
80	571	.85	.00	.01	.39	.80	.77	.92	6
101	297	1.00	.00	.05	.38	.81	.78	.90	6
105	358	1.27	.01	.07	.27	.83	.81	.91	5
132	329	.83	.00	.01	.35	.87	.77	.93	6
180	443	1.12	.00	.02	.30	.80	.82	.94	5
201	641	1.26	.00	.02	.27	.76	.85	.95	4
212	288	1.34	.00	.01	.24	.73	.87	.96	3
230	356	1.35	.00	.03	.23	.76	.86	.95	4
235	314	.99	.00	.02	.26	.71	.82	.93	5
301	591	.85	.00	.01	.33	.85	.78	.93	6
334	817	.90	.00	.03	.31	.81	.79	.92	6
341	448	.93	.00	.02	.29	.70	.81	.93	5
343	682	.97	.00	.01	.29	.78	.81	.94	5
346	692	.81	.00	.02	.33	.72	.78	.92	6
391	469	.58	.00	.02	.34	.81	.70	.88	10
501	454	.98	.01	.01	.28	.79	.80	.93	5
510	482	.96	.00	.01	.29	.75	.81	.94	5

511	680	1.14	.00	.02	.28	.70	.84	.94	4
560	486	.99	.00	.01	.30	.72	.82	.94	5
570	333	1.29	.00	.01	.21	.61	.88	.97	3
801	784	.80	.00	.03	.28	.71	.78	.91	6
950	347	1.21	.03	.05	.29	.88	.79	.90	6
996	518	1.29	.00	.06	.31	.92	.81	.92	5
1001	256	.61	.00	.02	.41	1.01	.66	.86	12
1035	419	1.08	.00	.01	.27	.73	.84	.95	4
1082	350	1.06	.00	.02	.29	.94	.80	.93	5
1083	361	1.08	.00	.02	.33	.81	.82	.94	5
1101	539	.76	.00	.04	.36	.79	.74	.89	8
1102	710	1.09	.00	.01	.29	.73	.84	.95	4
1150	422	.81	.00	.01	.36	.73	.78	.93	6
1165	662	.94	.00	.01	.33	.75	.81	.93	5
1170	512	.76	.00	.02	.32	.74	.77	.92	7
1301	531	.71	.00	.02	.29	.76	.75	.90	7
1701	284	.89	.00	.02	.33	.83	.78	.92	6
1801	369	.72	.00	.05	.36	.92	.70	.87	10
1810	450	1.02	.00	.04	.28	.92	.78	.91	6
1811	408	.39	.00	.01	.36	.76	.63	.86	13
1910	756	.57	.00	.03	.37	.75	.70	.88	10
2003	482	.90	.00	.02	.32	.82	.78	.92	6
2010	413	.88	.00	.02	.33	.82	.79	.93	6

2101	332	.43	.01	.04	.45	.81	.59	.79	22
2130	259	.81	.00	.02	.36	.78	.76	.91	7

Note. *n* = number of raters. i = item. g = grade. r = rater. irg,e = item × rater(grade) and error. Φs are calculated assuming either 5 or 20 raters. The *n* needed to obtain Φ = .80 was calculated by rounding up the estimate. All estimates assume a single grade and were calculated after setting negative variance components to 0.

Table 18

Generalizability analysis for item × rater(grade× tenure) design: PA.

		Variance Components							Φ		*n* for
Series	*n*	i	g	t	ig	it	r(gt)	irgt,e	*n* = 5	*n* = 20	Φ = .80
28	576	.74	.01	.01	.04	.00	.31	.77	.73	.81	9
201	635	1.32	.00	-.00	.02	.00	.27	.76	.85	.91	4
301	582	.87	-.00	.00	.01	.00	.33	.85	.78	.87	6
341	445	.91	.01	.01	.02	.00	.28	.70	.80	.87	6
343	672	.95	-.00	-.00	.01	.00	.29	.78	.81	.89	5
346	688	.83	.00	-.00	.02	-.00	.33	.72	.78	.87	6
1101	536	.79	-.00	.00	.04	.00	.35	.79	.75	.84	8
1301	530	.73	-.00	-.00	.03	.01	.29	.76	.75	.84	8
2101	323	.41	.01	-.01	.04	.00	.46	.81	.57	.69	28

Note. *n* = number of raters. i = item. g = grade. t = tenure. r = rater. irgt,e = item × rater(grade × tenure) and error. Φs are calculated assuming either 5 or 20 raters. The *n* needed to obtain Φ = .80 was calculated by rounding up the estimate. All estimates assume a single grade and tenure and were calculated after setting negative variance components to 0.

Table 19

Agreement by tenure: CT.

	r_{wg}						a_{wg}					
	1-12 Mo.		13-48 Mo.		49+ Mo.		1-12 Mo.		13-48 Mo.		49+ Mo.	
Item	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	.58	.20	.54	.17	.53	.16	.72	.09	.68	.09	.67	.07
2	.51	.28	.48	.27	.51	.16	.70	.12	.67	.10	.68	.06
3	.52	.22	.48	.15	.52	.09	.72	.10	.68	.07	.70	.05
4	.59	.19	.53	.16	.53	.11	.76	.10	.72	.08	.72	.05
5	.35	.31	.33	.19	.35	.16	.67	.14	.64	.08	.65	.06
6	.49	.22	.49	.17	.51	.11	.71	.14	.71	.08	.71	.05
7	.42	.25	.49	.16	.49	.13	.69	.13	.70	.07	.70	.05
8	.34	.29	.40	.16	.40	.11	.69	.15	.70	.08	.70	.05
9	.26	.29	.29	.14	.30	.14	.66	.13	.65	.07	.65	.07
10	.57	.22	.58	.13	.59	.09	.73	.13	.74	.07	.73	.05
11	.25	.31	.22	.19	.30	.12	.64	.15	.62	.09	.65	.06
12	.39	.24	.35	.21	.38	.11	.71	.11	.68	.09	.69	.06
13	.31	.26	.28	.23	.30	.16	.64	.11	.60	.10	.61	.08
14	.25	.28	.27	.18	.31	.13	.63	.13	.63	.08	.64	.06
15	.46	.27	.39	.16	.39	.13	.72	.13	.68	.08	.67	.06
16	.59	.20	.62	.12	.61	.10	.73	.11	.71	.05	.71	.05
17	.44	.25	.47	.13	.46	.11	.70	.11	.70	.06	.70	.06
18	.47	.29	.50	.13	.53	.09	.70	.13	.70	.06	.71	.04

19	.27	.26	.25	.18	.30	.13	.65	.13	.63	.08	.65	.06
20	.25	.36	.25	.22	.31	.17	.63	.17	.62	.10	.63	.08
21	.68	.17	.66	.10	.65	.08	.75	.09	.73	.05	.73	.04
22	.68	.16	.68	.12	.68	.09	.71	.09	.70	.05	.69	.05
23	.66	.21	.66	.12	.66	.10	.75	.10	.72	.06	.72	.04
24	.57	.25	.59	.12	.61	.09	.72	.10	.72	.07	.73	.05
25	.65	.19	.62	.10	.64	.06	.76	.10	.74	.06	.75	.04
26	.60	.21	.62	.10	.62	.07	.74	.09	.73	.05	.74	.04
27	.34	.30	.37	.19	.41	.14	.68	.13	.67	.08	.68	.07
28	.39	.22	.43	.16	.46	.12	.69	.11	.68	.07	.69	.06
29	.51	.23	.52	.13	.55	.10	.71	.11	.70	.06	.71	.05
30	.42	.32	.39	.23	.43	.12	.71	.14	.68	.09	.69	.06
31	.58	.27	.63	.14	.64	.11	.66	.14	.67	.07	.68	.06

Table 20

Agreement by tenure: PA.

	r_{wg}						a_{wg}					
	1-12 Mo.		13-48 Mo.		49+ Mo.		1-12 Mo.		13-48 Mo.		49+ Mo.	
Item	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	.76	.24	.79	.13	.78	.11	.70	.17	.70	.11	.68	.07
2	.69	.24	.73	.16	.73	.13	.73	.14	.71	.10	.70	.07
3	.34	.30	.35	.22	.35	.18	.65	.14	.63	.09	.63	.06
4	.21	.33	.18	.24	.21	.17	.60	.16	.56	.10	.56	.07
5	.64	.24	.63	.20	.61	.18	.72	.13	.67	.10	.67	.08
6	.44	.29	.42	.19	.42	.12	.72	.14	.69	.09	.68	.05
7	.48	.27	.50	.19	.47	.15	.71	.12	.70	.10	.68	.07
8	.57	.21	.58	.17	.55	.15	.71	.12	.70	.09	.68	.08
9	.57	.24	.58	.19	.56	.14	.73	.12	.70	.10	.68	.07
10	.57	.26	.57	.18	.56	.16	.73	.13	.70	.10	.69	.07
11	.17	.35	.18	.23	.20	.17	.59	.17	.59	.11	.59	.08
12	.50	.21	.50	.14	.48	.11	.74	.10	.72	.07	.71	.06
13	.56	.24	.56	.17	.56	.11	.71	.12	.69	.08	.69	.06
14	.57	.26	.57	.18	.56	.13	.70	.14	.70	.08	.69	.06
15	.71	.24	.71	.16	.70	.12	.69	.15	.68	.10	.66	.07
16	.71	.15	.71	.14	.70	.08	.75	.10	.74	.08	.72	.05
17	.68	.18	.70	.16	.67	.11	.73	.09	.72	.09	.70	.06
18	.53	.26	.57	.16	.55	.14	.72	.13	.71	.08	.70	.07

19	.65	.23	.68	.12	.67	.08	.72	.12	.73	.07	.72	.05
20	.16	.40	.16	.33	.17	.26	.51	.18	.50	.15	.49	.10
21	-.03	.38	.03	.25	.03	.20	.46	.16	.47	.11	.47	.09
22	.13	.34	.15	.24	.19	.17	.58	.16	.57	.11	.58	.08
23	.26	.31	.29	.22	.29	.16	.64	.13	.64	.10	.63	.07
24	.34	.45	.33	.32	.30	.28	.58	.19	.57	.13	.56	.10
25	.52	.26	.48	.18	.46	.12	.73	.13	.70	.08	.68	.05
26	.31	.30	.31	.21	.31	.15	.66	.14	.66	.10	.65	.07
27	.32	.35	.37	.19	.36	.13	.67	.15	.68	.09	.67	.07
28	.27	.32	.30	.22	.29	.17	.65	.14	.64	.10	.63	.08
29	.09	.37	.11	.27	.13	.19	.55	.17	.56	.12	.55	.09
30	.48	.26	.50	.18	.50	.13	.70	.13	.69	.09	.68	.06
31	.57	.25	.59	.14	.58	.09	.74	.12	.74	.07	.73	.05
32	.48	.30	.47	.20	.46	.14	.70	.14	.69	.10	.67	.07
33	.46	.35	.43	.25	.47	.17	.66	.16	.64	.11	.65	.07
34	.56	.22	.58	.14	.57	.08	.76	.11	.75	.07	.74	.06
35	.37	.29	.37	.20	.35	.14	.70	.13	.69	.10	.67	.07
36	.38	.42	.42	.26	.42	.20	.60	.18	.58	.12	.57	.07
37	.27	.37	.35	.22	.32	.17	.58	.17	.59	.11	.57	.07
38	.62	.33	.64	.23	.61	.21	.64	.21	.58	.15	.56	.08
39	.29	.36	.37	.24	.31	.22	.56	.16	.57	.11	.54	.08
40	.32	.44	.41	.28	.38	.24	.55	.18	.56	.14	.55	.08
41	.27	.34	.31	.32	.29	.23	.54	.16	.53	.13	.50	.08

42	.45	.40	.49	.27	.43	.23	.54	.22	.55	.14	.51	.10
43	.49	.39	.51	.27	.46	.25	.55	.21	.55	.15	.52	.10
44	.41	.43	.47	.33	.41	.27	.54	.21	.54	.15	.52	.11

Table 21

Regression of agreement on occupational tenure: CT.

	r_{wg}							a_{wg}						
			b							*b*				
Item	R^2	ΔR^2	*a*	*M*	M^2	D1	D2	R^2	ΔR^2	*a*	*M*	M^2	D1	D2
1	.52*	.00	.54	.37*	.27*	-.03	-.02	.12*	.05*	.70	.04*	.18*	-.04*	-.04*
2	.42*	.00	.46	.39*	.27*	-.02	.01	.09*	.02	.68	.06*	.12*	-.03	-.02
3	.24*	.02	.50	.27*	.14	-.03	.01	.01	.04*	.72	.00	.04	-.04*	-.01
4	.24*	.01	.57	.24*	.12	-.05	-.03	.01	.05*	.76	.01	.02	-.04*	-.04*
5	.29*	.00	.27	.14*	.24*	.00	.02	.02	.01	.65	.00	.04	-.02	-.02
6	.08*	.01	.46	.14*	.23	.01	.04	.02	.00	.71	-.04	.05	.00	.00
7	.22*	.02	.40	.24*	.21	.06	.06	.03	.01	.68	.02	.09	.02	.01
8	.01	.02	.32	.01	.13	.06	.07	.01	.00	.69	-.03	-.01	.01	.02
9	.02	.01	.23	.04	.26	.05	.05	.02	.00	.65	.04	.06	.00	.00
10	.11*	.00	.56	.23*	.06	.02	.02	.01	.00	.73	-.03	.02	.00	.00
11	.02	.03	.22	.04	.13	-.01	.07	.00	.02	.64	-.02	-.01	-.02	.01

12	.01	.01	.37	.04	.10	-.03	.01	.00	.01	.71	-.02	.00	-.03	-.02
13	.28*	.01	.25	-.25*	.24*	-.03	.01	.01	.02	.63	-.01	.03	-.04	-.03
14	.10*	.02	.22	.17*	.12	.04	.07	.01	.00	.63	.02	-.01	.00	.01
15	.16*	.01	.42	.16*	.21	-.04	-.05	.01	.04*	.71	-.01	.05	-.04*	-.04*
16	.46*	.00	.62	.43*	.00	-.01	-.01	.02	.03	.74	.05	-.07	-.03	-.03
17	.17*	.01	.42	.22*	.15	.03	.03	.00	.00	.70	.01	.03	.00	-.01
18	.27*	.03*	.42	.33*	.35*	.05	.09*	.04	.01	.68	.04	.12*	.01	.02
19	.05	.01	.25	.12	.12	-.01	.04	.01	.01	.66	.00	-.05	-.02	-.01
20	.16*	.01	.20	.13*	.18*	.02	.07	.00	.00	.63	.00	.01	-.02	.00
21	.46*	.00	.65	.34*	.24*	-.01	.00	.02	.02	.74	.00	.11	-.02	-.02
22	.67*	.00	.67	.48*	.17	.01	.00	.06*	.02	.70	.07*	.19*	-.01	-.02
23	.46*	.00	.67	.38*	.00	-.01	-.01	.00	.04	.75	.00	-.09	-.03	-.03*
24	.29*	.01	.57	.36*	-.08	.03	.04	.01	.00	.73	.02	-.07	.00	.00
25	.24*	.01	.62	.24*	.23*	.00	.02	.03	.02	.76	-.04	.05	-.02	-.01
26	.39*	.00	.60	.35*	.02	.01	.02	.01	.00	.75	.00	-.07	-.01	-.01
27	.23*	.01	.31	.21*	.27*	.04	.06	.07*	.00	.66	.03	.11*	.00	.00

28	.23*	.02	.38	.19*	.23*	.02	.05	.02	.00	.68	.00	.07	-.01	.00
29	.30*	.01	.49	.31*	.28*	.02	.05	.03	.00	.70	.03	.10	.00	.01
30	.12*	.01	.38	.24*	.28*	-.02	.03	.04	.01	.70	.04	.10	-.02	-.01
31	.64*	.02*	.56	.57*	.25*	.05*	.07*	.04*	.01	.64	.18*	.22*	.02	.03

Note. R^2 is for step 1, which includes M and M^2. ΔR^2 is the change after adding D1 and D2. *b* is the unstandardized regression coefficient, calculated from the final model. D1 and D2 represent tenure of 12 months or less vs. 13-48 months and vs. 49 or more months, respectively.

* $p < .05$.

Table 22

Regression of agreement on occupational tenure: PA.

	r_{wg}							a_{wg}						
			b							*b*				
Item	R^2	ΔR^2	*a*	*M*	M^2	D1	D2	R^2	ΔR^2	*a*	*M*	M^2	D1	D2
1	.67*	.00	.80	.60*	-.39*	-.01	-.01	.13*	.02	.72	.25*	.20	-.03	-.05
2	.64*	.00	.70	.48*	.12	.01	.02	.04*	.02	.73	.08*	.09	-.02	-.03
3	.43*	.00	.26	.20*	.20*	.00	.02	.00	.01	.65	.01	.01	-.02	-.02
4	.31*	.00	.10	-.06*	.25*	-.03	.00	.03*	.02*	.59	.02	.02	-.04*	-.04*
5	.62*	.00	.63	.48*	.13*	-.03	-.02	.03	.05*	.72	.07*	.02	-.06*	-.06*
6	.13*	.00	.42	.17*	.10*	-.02	-.02	.01	.02	.72	.00	-.03	-.03	-.04*
7	.26*	.00	.49	.35*	-.01	.00	-.01	.03*	.02	.72	.05	-.06	-.03	-.04*
8	.43*	.00	.56	.38*	.16*	.00	-.02	.00	.02	.72	.01	.00	-.02	-.04*
9	.40*	.00	.58	.38*	.03	-.02	-.02	.01	.03*	.73	.04	.01	-.03	-.05*
10	.35*	.00	.58	.42*	.03	-.01	-.02	.02	.03*	.73	.04	-.05	-.03	-.04*
11	.10*	.01	.09	.06	.23*	.04	.07	.06*	.00	.58	.06*	.04	.00	.00

12	.11*	.00	.51	.18*	.01	-.01	-.02	.01	.03*	.75	-.02	-.06	-.02	-.03*
13	.41*	.00	.55	.41*	.07	.00	.01	.02	.01	.71	.05*	.03	-.02	-.02
14	.39*	.00	.53	.44*	.16*	.02	.03	.02	.00	.70	.05	.02	.00	-.01
15	.62*	.00	.69	.60*	.10	.01	.01	.17*	.01	.67	.19*	.48*	-.01	-.03
16	.37*	.00	.71	.39*	.06	-.01	-.01	.00	.02	.75	.00	.00	-.01	-.03
17	.58*	.00	.68	.44*	.01	.01	.00	.01	.03*	.74	.02	-.09	-.02	-.04*
18	.31*	.00	.54	.42*	.07	.02	.01	.03*	.01	.72	.07*	.04	-.01	-.02
19	.36*	.01	.63	.35*	.19	.03	.03	.01	.00	.72	-.02	.10	.01	.00
20	.49*	.00	.01	-.28*	.26*	.03	.05	.02	.00	.50	.03	.00	-.01	-.02
21	.28*	.04*	-.16	-.21*	.33*	.10*	.14*	.05*	.00	.45	.04*	.04	.01	.01
22	.19*	.02	.06	.19*	.23*	.04	.08*	.09*	.00	.58	.07*	.00	.00	.00
23	.13*	.01	.22	.14*	.17*	.04	.05	.01	.00	.65	.02	-.01	.00	-.01
24	.58*	.00	.24	.56*	.21*	.03	.05	.14*	.00	.58	.11*	.02	-.02	-.02
25	.21*	.00	.48	.25*	.20*	-.03	-.03	.01	.03*	.72	.00	.03	-.03*	-.04*
26	.05*	.00	.26	.03	.13*	.03	.03	.01	.01	.67	.02	-.04	-.01	-.02
27	.09*	.02	.28	.15*	.17*	.07	.07	.00	.00	.67	.02	.01	.01	.00

28	.18*	.00	.22	.19*	.16*	.04	.04	.04*	.01	.65	.05*	.00	-.01	-.02
29	.09*	.01	.01	.06	.21*	.06	.08	.02	.00	.56	.03	-.03	.00	.00
30	.28*	.00	.46	.29*	.17*	.02	.02	.00	.00	.70	.00	.02	-.01	-.01
31	.22*	.01	.55	.29*	.09	.03	.03	.00	.00	.74	.00	-.05	.00	-.01
32	.27*	.00	.45	.33*	.13	.01	.01	.02	.01	.70	.04	.03	-.02	-.03
33	.42*	.00	.44	.53*	.04	.00	.02	.07*	.00	.67	.10*	-.01	-.02	-.02
34	.11*	.00	.55	.18*	.10	.02	.01	.02	.01	.76	-.04	-.04	-.01	-.02
35	.03*	.00	.38	.11*	-.01	-.01	-.03	.01	.02	.71	.02	-.07	-.02	-.04*
36	.53*	.00	.30	-.51*	.30*	.04	.05	.07*	.01	.58	-.06*	.08*	-.02	-.03
37	.46*	.01*	.21	-.39*	.32*	.07*	.07*	.05*	.00	.56	-.01	.08*	.01	.00
38	.73*	.00	.57	-.69*	.28*	.01	.00	.10*	.04*	.63	-.16*	.12*	-.06*	-.08*
39	.54*	.01	.23	-.49*	.28*	.05	.03	.04*	.02	.55	-.05*	.05*	.01	-.02
40	.61*	.01*	.22	-.54*	.31*	.08*	.09*	.06*	.00	.53	-.06*	.08*	.01	.00
41	.62*	.00	.18	-.52*	.32*	.05	.04	.04*	.02	.52	-.04	.06*	.00	-.03
42	.69*	.00	.36	-.68*	.33*	.04	.02	.12*	.02	.52	-.12*	.11*	.01	-.03
43	.71*	.00	.39	-.69*	.34*	.03	.02	.13*	.01	.53	-.13*	.13*	.00	-.04

44	.72*	.00	.30	-.67*	.32*	.05	.05	.12*	.00	.51	-.12*	.12*	.00	-.02

Note. R^2 is for step 1, which includes M and M^2. ΔR^2 is the change after adding D1 and D2. *b* is the unstandardized regression coefficient, calculated from the final model. D1 and D2 represent tenure of 12 months or less vs. 13 – 48 months and vs. 49 or more months, respectively.

* $p < .05$.

Table 23

Agreement by occupational complexity: CT.

	r_{wg}										a_{wg}									
	4		5		6		7		8		4		5		6		7		8	
Item	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	.44	.15	.49	.11	.57	.10	.57	.10	.61	.14	.65	.05	.66	.03	.69	.04	.68	.02	.68	.04
2	.40	.14	.45	.10	.52	.11	.54	.19	.59	.17	.65	.05	.66	.03	.69	.04	.67	.08	.69	.06
3	.48	.10	.47	.09	.51	.07	.51	.09	.55	.10	.69	.05	.69	.04	.70	.04	.69	.04	.71	.05
4	.50	.09	.51	.07	.53	.08	.55	.10	.58	.13	.71	.04	.72	.03	.72	.03	.72	.04	.73	.05
5	.30	.16	.30	.12	.36	.14	.39	.13	.38	.15	.64	.06	.62	.03	.64	.05	.66	.05	.65	.05
6	.45	.10	.46	.07	.50	.08	.54	.08	.56	.09	.71	.05	.70	.03	.70	.03	.71	.04	.72	.04
7	.41	.11	.46	.07	.48	.11	.52	.12	.57	.11	.68	.06	.69	.03	.69	.05	.70	.05	.72	.04
8	.36	.12	.38	.10	.40	.06	.40	.08	.43	.08	.68	.06	.69	.05	.70	.03	.70	.04	.71	.03
9	.26	.10	.31	.08	.30	.06	.32	.08	.33	.10	.63	.05	.65	.04	.65	.03	.66	.04	.67	.05
10	.54	.10	.57	.07	.57	.06	.61	.04	.62	.04	.71	.06	.73	.04	.72	.03	.74	.02	.74	.03
11	.29	.09	.28	.10	.25	.09	.29	.09	.26	.11	.64	.05	.63	.05	.62	.05	.64	.04	.63	.05

12	.35	.13	.38	.07	.39	.07	.38	.08	.35	.09	.67	.07	.69	.04	.70	.03	.69	.04	.68	.05
13	.22	.10	.29	.11	.33	.17	.32	.13	.35	.13	.59	.08	.61	.05	.62	.06	.60	.04	.62	.06
14	.32	.11	.34	.08	.28	.04	.25	.05	.23	.11	.64	.06	.65	.04	.63	.03	.62	.03	.62	.06
15	.40	.10	.41	.09	.39	.09	.40	.07	.37	.12	.67	.05	.68	.04	.67	.05	.68	.04	.66	.05
16	.55	.09	.60	.06	.63	.08	.64	.05	.67	.06	.70	.05	.71	.03	.71	.04	.71	.03	.71	.03
17	.43	.13	.46	.07	.46	.07	.46	.12	.50	.09	.69	.06	.70	.03	.69	.04	.68	.06	.69	.03
18	.46	.08	.49	.09	.53	.08	.53	.08	.56	.07	.70	.04	.69	.04	.71	.04	.70	.03	.70	.03
19	.31	.07	.33	.11	.29	.08	.26	.06	.26	.13	.65	.04	.66	.05	.64	.04	.62	.03	.62	.06
20	.27	.15	.24	.14	.25	.11	.25	.15	.34	.13	.63	.08	.62	.07	.61	.05	.60	.07	.60	.09
21	.61	.05	.64	.06	.65	.06	.68	.05	.69	.09	.72	.03	.73	.03	.72	.03	.73	.03	.73	.03
22	.64	.09	.65	.05	.67	.07	.70	.07	.75	.07	.69	.04	.68	.03	.68	.04	.68	.03	.71	.03
23	.63	.09	.63	.09	.65	.09	.67	.08	.71	.09	.71	.04	.71	.03	.71	.03	.72	.04	.73	.05
24	.56	.13	.58	.06	.59	.07	.61	.09	.64	.06	.71	.07	.72	.04	.72	.03	.72	.05	.73	.02
25	.60	.06	.62	.05	.63	.05	.64	.04	.65	.05	.74	.04	.75	.04	.74	.03	.75	.02	.73	.04
26	.60	.06	.60	.04	.60	.07	.63	.03	.66	.07	.74	.05	.74	.02	.72	.03	.73	.02	.73	.04
27	.35	.10	.37	.08	.38	.11	.40	.15	.48	.13	.67	.05	.68	.04	.67	.04	.67	.07	.68	.05

28	.41	.08	.39	.07	.45	.07	.48	.11	.53	.13	.68	.04	.67	.04	.69	.03	.69	.05	.69	.04
29	.52	.10	.52	.06	.54	.08	.53	.08	.59	.10	.70	.06	.70	.03	.70	.03	.70	.03	.71	.04
30	.39	.11	.41	.07	.43	.07	.43	.11	.47	.13	.68	.06	.69	.03	.69	.03	.68	.04	.70	.07
31	.61	.13	.62	.10	.63	.10	.67	.10	.65	.11	.66	.05	.67	.04	.68	.04	.68	.03	.66	.06

Table 24

Agreement by complexity level: PA.

	r_{wg}								a_{wg}							
	9		11		12		13		9		11		12		13	
Item	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	.74	.08	.77	.08	.80	.09	.83	.10	.67	.06	.68	.05	.68	.05	.68	.08
2	.66	.13	.71	.10	.75	.11	.80	.09	.69	.06	.69	.05	.70	.04	.70	.08
3	.37	.16	.36	.16	.35	.17	.33	.16	.63	.04	.62	.04	.62	.05	.63	.06
4	.19	.14	.20	.14	.20	.18	.22	.18	.55	.05	.55	.06	.55	.06	.57	.06
5	.56	.15	.61	.15	.63	.18	.69	.13	.65	.08	.66	.06	.66	.06	.68	.06
6	.38	.10	.40	.10	.42	.10	.48	.12	.68	.05	.68	.05	.68	.04	.69	.06
7	.45	.12	.48	.12	.48	.14	.49	.15	.68	.06	.69	.06	.68	.06	.69	.07
8	.50	.10	.54	.11	.58	.12	.62	.14	.67	.07	.68	.05	.68	.06	.69	.09
9	.51	.11	.54	.12	.58	.12	.63	.14	.68	.06	.68	.06	.68	.06	.70	.08
10	.51	.12	.56	.11	.56	.16	.62	.14	.68	.06	.69	.06	.68	.08	.69	.07
11	.19	.11	.20	.14	.20	.11	.22	.16	.57	.05	.58	.06	.58	.06	.60	.08
12	.50	.08	.50	.08	.48	.10	.47	.11	.71	.04	.71	.04	.71	.05	.71	.06
13	.58	.10	.56	.08	.53	.11	.54	.09	.69	.04	.68	.04	.67	.05	.68	.04
14	.56	.11	.56	.11	.54	.13	.58	.11	.68	.06	.68	.04	.68	.05	.70	.05
15	.70	.11	.71	.10	.69	.10	.71	.11	.66	.05	.66	.04	.65	.05	.66	.06
16	.70	.08	.70	.06	.69	.07	.71	.06	.73	.05	.73	.04	.72	.05	.73	.06
17	.69	.11	.68	.10	.65	.12	.69	.10	.70	.05	.70	.05	.69	.05	.71	.05
18	.53	.11	.56	.09	.56	.13	.58	.12	.69	.05	.70	.05	.70	.06	.71	.06

19	.66	.07	.68	.06	.66	.06	.67	.08	.70	.04	.72	.03	.72	.04	.74	.05
20	.11	.23	.15	.22	.20	.24	.24	.25	.45	.09	.47	.07	.49	.08	.54	.11
21	-.01	.17	.00	.15	.07	.17	.09	.18	.43	.08	.45	.06	.48	.07	.52	.08
22	.13	.15	.17	.12	.19	.12	.25	.19	.55	.07	.57	.05	.58	.06	.60	.09
23	.25	.17	.30	.12	.29	.10	.35	.12	.61	.06	.63	.05	.63	.05	.66	.06
24	.33	.26	.33	.24	.29	.27	.30	.22	.55	.10	.56	.08	.55	.09	.58	.08
25	.45	.10	.47	.09	.46	.11	.49	.10	.68	.05	.69	.04	.68	.05	.70	.05
26	.27	.10	.30	.09	.32	.09	.36	.12	.63	.05	.64	.05	.65	.05	.67	.06
27	.32	.11	.36	.09	.36	.10	.41	.10	.65	.05	.67	.04	.67	.05	.69	.04
28	.24	.13	.24	.12	.30	.12	.38	.14	.61	.07	.61	.06	.63	.05	.67	.07
29	.08	.13	.10	.11	.14	.11	.23	.13	.53	.07	.54	.05	.55	.06	.60	.07
30	.50	.10	.51	.08	.49	.11	.50	.11	.67	.05	.68	.04	.68	.05	.69	.05
31	.57	.08	.58	.05	.58	.05	.59	.08	.72	.05	.73	.03	.73	.03	.73	.05
32	.47	.12	.48	.11	.45	.11	.45	.13	.67	.06	.67	.05	.67	.05	.68	.07
33	.45	.15	.48	.13	.47	.16	.46	.17	.65	.06	.65	.05	.64	.07	.64	.08
34	.57	.09	.57	.05	.56	.04	.58	.06	.74	.05	.73	.04	.73	.03	.75	.04
35	.34	.13	.36	.09	.34	.12	.37	.14	.67	.06	.67	.04	.66	.06	.68	.07
36	.32	.17	.38	.15	.46	.13	.53	.18	.56	.06	.58	.06	.56	.05	.56	.07
37	.25	.12	.29	.14	.36	.13	.41	.16	.56	.05	.57	.06	.56	.05	.57	.06
38	.53	.17	.56	.17	.67	.16	.74	.18	.55	.06	.55	.07	.57	.07	.58	.08
39	.24	.15	.29	.17	.36	.17	.42	.18	.54	.07	.54	.06	.54	.06	.55	.07
40	.30	.19	.37	.21	.39	.21	.50	.21	.52	.07	.55	.06	.53	.06	.56	.06
41	.22	.20	.25	.18	.33	.21	.38	.21	.50	.07	.50	.06	.50	.06	.51	.08

42	.35	.19	.40	.19	.47	.17	.58	.20	.51	.08	.52	.09	.49	.08	.51	.08
43	.39	.19	.43	.19	.50	.20	.60	.22	.51	.08	.51	.06	.50	.08	.53	.12
44	.35	.24	.38	.25	.46	.27	.53	.26	.50	.08	.51	.10	.51	.08	.54	.12

Table 25

Mean importance by occupational complexity: CT.

	4		5		6		7		8		
Item	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*
1	3.91	1.08	4.05	1.01	4.18	.93	4.17	.95	4.29	.91	43.20*
2	3.80	1.14	3.92	1.05	4.04	1.00	4.11	.97	4.21	.94	52.89*
3	3.79	1.04	3.81	1.04	3.90	1.00	3.94	1.00	3.99	.96	17.51*
4	3.73	1.02	3.80	.99	3.86	.97	3.94	.94	3.99	.92	27.55*
5	3.19	1.24	3.41	1.24	3.45	1.24	3.49	1.22	3.41	1.25	17.42*
6	3.49	1.05	3.62	1.05	3.73	1.03	3.86	.98	3.93	.99	68.15*
7	3.52	1.10	3.61	1.06	3.72	1.05	3.87	1.00	3.99	.96	76.68*
8	3.05	1.14	3.12	1.12	3.17	1.09	3.27	1.10	3.30	1.09	20.81*
9	2.85	1.21	2.81	1.18	2.83	1.19	2.92	1.18	2.98	1.19	7.27*
10	3.88	.97	3.91	.94	3.97	.93	3.98	.89	4.07	.88	12.92*
11	3.37	1.18	3.28	1.21	3.21	1.24	3.15	1.21	3.11	1.23	14.94*
12	3.28	1.14	3.18	1.12	3.20	1.10	3.15	1.11	3.14	1.14	4.87*
13	2.77	1.30	2.46	1.22	2.36	1.21	2.31	1.21	2.29	1.20	42.11*
14	3.52	1.17	3.39	1.17	3.33	1.21	3.16	1.22	3.05	1.25	51.75*
15	3.54	1.12	3.54	1.10	3.56	1.11	3.54	1.11	3.55	1.12	0.15
16	3.96	.99	4.08	.91	4.17	.89	4.23	.86	4.28	.84	43.59*
17	3.58	1.09	3.66	1.04	3.70	1.05	3.75	1.04	3.84	1.03	18.83*
18	3.68	1.06	3.82	1.01	3.89	.98	3.93	.99	4.01	.97	30.78*
19	3.31	1.19	3.31	1.17	3.30	1.19	3.28	1.22	3.27	1.24	0.37

20	3.25	1.24	3.20	1.24	3.25	1.25	3.31	1.25	3.64	1.21	31.46*
21	4.10	.90	4.19	.85	4.26	.83	4.27	.80	4.37	.77	31.40*
22	4.30	.86	4.36	.83	4.43	.80	4.46	.77	4.53	.71	27.26*
23	4.20	.87	4.24	.85	4.28	.83	4.30	.81	4.35	.77	10.87*
24	4.01	.94	4.06	.90	4.10	.90	4.12	.87	4.19	.85	11.37*
25	3.96	.90	4.01	.88	4.08	.87	4.11	.85	4.17	.83	21.21*
26	3.99	.92	4.02	.89	4.07	.89	4.12	.86	4.20	.84	20.02*
27	3.38	1.15	3.39	1.12	3.43	1.11	3.53	1.10	3.77	1.07	41.99*
28	3.55	1.10	3.53	1.12	3.60	1.07	3.69	1.05	3.91	1.03	42.63*
29	3.86	1.02	3.89	1.00	3.93	.98	3.95	.98	4.05	.93	10.84*
30	3.49	1.11	3.48	1.10	3.51	1.08	3.58	1.08	3.61	1.08	6.15*
31	4.27	.93	4.29	.91	4.29	.90	4.36	.85	4.37	.86	7.34*

* $p < .05$.

Table 26

Mean importance by occupational complexity: PA.

	9		11		12		13		
Item	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*
1	4.54	.73	4.60	.69	4.66	.64	4.71	.59	58.77*
2	4.31	.86	4.41	.81	4.52	.73	4.60	.66	137.28*
3	3.59	1.25	3.55	1.27	3.56	1.27	3.57	1.25	0.87
4	2.77	1.39	2.81	1.42	2.88	1.42	2.98	1.40	21.22*
5	4.15	.99	4.28	.92	4.31	.92	4.47	.78	98.25*
6	3.25	1.16	3.39	1.13	3.55	1.12	3.81	1.03	225.44*
7	3.78	1.06	3.84	1.02	3.84	1.03	3.87	.99	6.84*
8	3.95	1.03	4.05	.98	4.18	.92	4.27	.85	108.88*
9	3.93	1.02	4.03	.98	4.17	.92	4.26	.85	115.59*
10	3.96	1.00	4.07	.95	4.16	.93	4.27	.85	92.20*
11	2.79	1.33	2.87	1.34	2.91	1.32	3.04	1.28	29.49*
12	3.69	1.02	3.70	1.01	3.67	1.02	3.65	1.02	3.34
13	4.14	.94	4.10	.95	4.05	.98	4.04	.96	11.93*
14	4.12	.95	4.08	.96	4.05	.98	4.12	.92	6.98*
15	4.48	.81	4.50	.78	4.49	.79	4.53	.76	3.36
16	4.33	.80	4.35	.78	4.34	.79	4.35	.76	0.49
17	4.36	.82	4.36	.82	4.31	.85	4.34	.81	5.44*
18	3.98	.98	4.04	.95	4.02	.95	4.08	.90	8.89*
19	4.29	.84	4.29	.81	4.24	.83	4.18	.83	20.42*

20	2.39	1.46	2.37	1.44	2.41	1.43	2.59	1.39	25.00*
21	2.45	1.46	2.47	1.45	2.49	1.41	2.79	1.38	60.80*
22	3.07	1.36	3.11	1.34	3.17	1.32	3.48	1.22	96.30*
23	3.30	1.27	3.26	1.24	3.25	1.21	3.34	1.16	5.67*
24	4.00	1.17	3.96	1.21	3.86	1.25	3.74	1.26	44.60*
25	3.78	1.06	3.77	1.05	3.71	1.06	3.76	1.03	6.20*
26	2.82	1.24	2.86	1.21	2.90	1.20	3.06	1.16	35.56*
27	3.33	1.18	3.31	1.15	3.30	1.15	3.42	1.11	12.59*
28	3.12	1.27	3.21	1.27	3.33	1.24	3.56	1.12	116.48*
29	2.91	1.40	2.99	1.39	3.02	1.36	3.18	1.26	32.53*
30	3.92	1.03	3.91	1.01	3.86	1.03	3.87	1.00	4.65
31	3.93	.93	3.94	.93	3.93	.92	3.95	.91	0.38
32	3.90	1.05	3.87	1.05	3.85	1.05	3.74	1.06	21.29*
33	3.95	1.07	3.99	1.04	4.04	1.03	4.04	1.03	8.42*
34	3.87	.93	3.85	.94	3.83	.94	3.83	.92	2.16
35	3.21	1.16	3.21	1.14	3.27	1.15	3.30	1.11	7.75*
36	2.17	1.22	2.04	1.18	1.82	1.09	1.71	1.04	171.01*
37	2.33	1.26	2.22	1.23	2.01	1.18	1.90	1.12	132.40*
38	1.68	1.03	1.60	1.00	1.45	.87	1.38	.80	113.41*
39	2.25	1.27	2.12	1.25	1.94	1.18	1.80	1.10	131.14*
40	2.06	1.26	2.05	1.24	1.89	1.17	1.80	1.11	58.24*
41	2.15	1.30	2.09	1.29	1.91	1.21	1.89	1.20	57.79*
42	1.90	1.19	1.83	1.17	1.65	1.08	1.56	1.00	101.05*

43	1.83	1.16	1.80	1.15	1.65	1.07	1.55	.99	75.66*
44	1.92	1.23	1.88	1.22	1.73	1.13	1.65	1.05	64.39*

Table 27

Regression of agreement on occupational complexity: CT.

	r_{wg}									a_{wg}								
			b									*b*						
Item	R^2	ΔR^2	*a*	*M*	M^2	D1	D2	D3	D4	R^2	ΔR^2	*a*	*M*	M^2	D1	D2	D3	D4
1	.77*	.02	.49	.39*	.07	.02	.06*	.05	.05	.07	.08	.66	.02	-.04	.01	.03*	.02	.02
2	.67*	.02	.43	.37*	.25	.04	.07*	.03	.05	.08	.06	.64	.03	.06	.02	.04	.02	.03
3	.48*	.01	.48	.29*	.32	.00	.02	.01	.02	.04	.01	.69	.02	.12	.00	.01	.00	.01
4	.54*	.00	.53	.31*	-.01	.01	.01	.01	.02	.08	.01	.72	.04	-.04	.00	.00	.00	.01
5	.57*	.03	.30	.17*	.15*	-.03	.02	.05	.04	.02	.08	.64	.02	-.01	-.02	.00	.02	.01
6	.46*	.01	.49	.24*	.23	-.01	-.01	.02	.01	.02	.02	.71	.01	.05	-.01	-.01	.00	.00
7	.60*	.01	.47	.34*	.24	.02	.00	.00	.00	.16*	.01	.69	.06	.09	.01	.00	.00	.00
8	.07	.03	.36	.05	.19	.02	.04	.03	.05	.01	.03	.68	.00	.05	.01	.02	.01	.02
9	.00	.08	.26	.00	-.06	.05	.04	.06	.07	.03	.06	.64	.02	-.11	.02	.02	.03	.03*
10	.23*	.10*	.53	.18*	.69*	.04	.03	.07*	.05	.06	.12*	.70	-.04	.36*	.02	.01	.04*	.03
11	.07	.04	.27	-.04	.40	-.01	-.05	-.01	-.04	.07	.04	.64	-.05	.13	-.01	-.03	-.01	-.02

12	.02	.04	.36	-.08	.04	.03	.04	.02	-.01	.06	.03	.68	-.08	-.02	.01	.02	.01	.00
13	.45*	.02	.26	-.23*	.07	.04	.05	.02	.05	.05	.04	.60	.00	-.06	.02	.03	.01	.03
14	.17*	.06	.30	.09	-.06	.03	-.02	-.04	-.05	.07	.06	.65	-.01	-.12	.01	-.01	-.02	-.02
15	.25*	.02	.39	.18*	.14	.02	.00	.01	-.02	.00	.02	.67	-.01	-.01	.00	.00	.00	-.01
16	.65*	.01	.61	.40*	.04	.02	.01	.00	.01	.05	.01	.71	.05	-.01	.01	.01	.00	.00
17	.32*	.01	.45	.24*	.10	.02	.00	-.01	.00	.00	.02	.69	.02	-.01	.01	.00	-.01	.00
18	.58*	.00	.52	.31*	.03	-.01	.01	.00	.00	.05	.01	.71	.03	-.05	-.01	.00	.00	.00
19	.10	.10	.31	.13	.16	.02	-.02	-.06	-.06	.01	.11	.65	.02	-.02	.00	-.01	-.03	-.03
20	.21*	.01	.27	.08	.17	-.02	-.03	-.04	-.01	.08	.02	.63	-.03	-.04	-.02	-.02	-.03	-.01
21	.59*	.01	.64	.34*	.42	.01	.00	.02	.00	.02	.02	.72	.00	.19	.00	.00	.01	.00
22	.81*	.01	.68	.43*	.09	.00	.00	.00	.02	.01	.09	.69	-.01	.00	-.01	-.01	.00	.02
23	.71*	.01	.66	.38*	-.02	-.01	.00	.00	.01	.03	.02	.72	.01	-.12	-.01	.00	.00	.01
24	.35*	.01	.59	.30*	-.03	.01	.01	.02	.03	.00	.01	.71	.00	.00	.00	.00	.01	.02
25	.29*	.02	.61	.17*	.14	.01	.01	.02	.01	.10	.02	.74	-.07*	.01	.01	.00	.01	.00
26	.37*	.01	.61	.20*	.37	.01	-.01	.00	.00	.08	.01	.73	-.05	.17	.00	-.01	.00	.00
27	.43*	.01	.34	.09	.34*	.03	.02	.03	.04	.07	.01	.65	-.04	.11*	.01	.01	.01	.02

28	.60*	.03	.41	.17*	.42*	-.01	.02	.03	.00	.16*	.05	.67	-.03	.20*	-.01	.01	.02	.00
29	.58*	.00	.53	.32*	.25	.00	.01	.00	.01	.10	.00	.69	.02	.16	.00	.00	.00	.00
30	.31*	.02	.39	.17	.41	.02	.03	.00	.03	.06	.02	.67	.00	.18	.01	.01	.00	.02
31	.83*	.02	.62	.58*	.10	.02	.03	.02	-.01	.20*	.06	.66	.12*	.06	.01	.02	.01	-.01

Note. R^2 is for step 1, which includes M and M^2. ΔR^2 is the change after adding D1, D2, D3, and D4. b is the unstandardized regression coefficient, calculated from the final model. D1, D2, D3, and D4 represent grade 4 vs. grade 5 , grade 6, grade 7, and grade 8, respectively.

* $p < .05$.

Table 28

Regression of agreement on occupational complexity: PA.

	r_{wg}								a_{wg}							
			b								b					
Item	R^2	ΔR^2	a	M	M^2	D1	D2	D3	R^2	ΔR^2	a	M	M^2	D1	D2	D3
1	.79*	.00	.78	.56*	.35	.00	.00	.00	.04	.01	.67	.07	.40	.00	.01	.00
2	.84*	.00	.72	.45*	.13	.01	.01	.01	.03	.01	.69	-.04	-.13	.00	.01	.01
3	.70*	.00	.29	.20*	.19*	-.01	-.01	.00	.01	.01	.63	.01	.00	-.01	-.01	.00
4	.56*	.00	.10	-.08*	.25*	.00	-.01	.02	.04	.02	.55	.02	.01	-.01	-.01	.01
5	.80*	.00	.61	.48*	.02	.01	.01	.02	.08*	.02	.66	.02	-.07	.01	.01	.02
6	.35*	.01	.38	.15*	.17*	.01	.01	.03	.00	.01	.67	-.01	.01	.01	.00	.02
7	.49*	.00	.50	.48*	-.69*	.00	.00	.00	.22*	.00	.70	.12*	-.43*	.00	.00	.00
8	.56*	.00	.55	.38*	-.02	.01	.01	.01	.05	.02	.68	-.02	-.17*	.01	.01	.03
9	.51*	.00	.56	.36*	-.04	.00	.00	.02	.04	.02	.69	-.01	-.14	.00	.00	.02
10	.52*	.01	.58	.49*	-.07	-.01	-.03	-.03	.09*	.01	.70	.09	-.12	.00	-.02	-.01
11	.22*	.01	.15	.06	.25*	.00	.01	.03	.18*	.01	.58	.07*	.02	.00	.00	.02

12	.21*	.00	.50	.25*	-.33	.00	-.02	-.01	.08*	.01	.72	.02	-.28*	.00	-.01	.00
13	.60*	.01	.56	.37*	.04	-.01	-.02	.00	.02	.02	.69	.01	-.07	-.01	-.02	.00
14	.61*	.01	.54	.48*	.20*	.01	.01	.03	.10*	.02	.68	.09*	.03	.01	.00	.02
15	.81*	.00	.70	.59*	.13	.01	.00	-.01	.09*	.00	.65	.09*	.16	.00	.00	.00
16	.36*	.01	.70	.33*	.26	.00	-.01	.00	.01	.01	.73	-.04	.05	.00	-.01	.00
17	.75*	.01	.68	.42*	.02	.00	-.01	.01	.02	.03	.70	.00	-.11	.00	-.01	.01
18	.56*	.00	.56	.46*	-.49*	.00	.01	.02	.18*	.01	.71	.08*	-.36*	.00	.01	.01
19	.43*	.03	.65	.27*	.12	.02	.01	.03*	.11*	.04	.71	-.08*	-.12	.01	.01	.03*
20	.59*	.05*	.01	-.27*	.23*	.04	.08*	.15*	.09*	.11*	.46	.05*	-.02	.02	.04	.08*
21	.31*	.07*	-.07	-.14*	.28*	.03	.08*	.12*	.19*	.09*	.44	.07*	-.01	.01	.04*	.06*
22	.38*	.01	.13	.22*	.21*	.02	.03	.03	.23*	.01	.57	.08*	-.01	.01	.01	.02
23	.25*	.07*	.23	.16*	.11	.05	.05	.10*	.04	.08*	.61	.02	-.04	.02	.02	.05*
24	.73*	.01	.24	.54*	.23*	.01	.01	.06	.22*	.03	.54	.10*	.02	.01	.01	.04
25	.32*	.01	.44	.23*	.24*	.02	.02	.03	.02	.02	.67	.00	.05	.01	.01	.02
26	.06	.07*	.27	.02	.08	.03	.05	.08*	.06	.07*	.64	.02	-.05	.01	.02	.04*
27	.18*	.06*	.31	.13*	.14	.04	.04	.07*	.01	.08*	.65	.01	-.01	.02	.02	.04*

28	.32*	.04	.26	.18*	.07	-.02	.02	.06	.12*	.07*	.63	.03	-.06	-.01	.01	.04
29	.13*	.14*	.07	.09*	.10	.02	.05	.13*	.15*	.14*	.54	.04*	-.08*	.01	.02	.07*
30	.44*	.01	.49	.28*	.01	.01	.00	.02	.03	.02	.68	-.01	-.08	.01	.00	.02
31	.18*	.01	.57	.18*	-.17	.01	.01	.02	.09*	.01	.73	-.06*	-.21	.01	.01	.01
32	.38*	.01	.45	.29*	.04	.01	.00	.02	.02	.01	.67	.00	-.08	.00	.00	.01
33	.62*	.01	.47	.48*	.01	.02	-.01	-.01	.11*	.01	.65	.08*	-.06	.01	-.01	-.01
34	.09*	.01	.56	.09	.34	.00	.00	.02	.12*	.01	.73	-.08*	.09	.00	.00	.01
35	.16*	.01	.37	.18*	-.44	.02	-.01	.01	.12*	.01	.68	.05	-.30*	.01	-.01	.01
36	.80*	.00	.36	-.46*	.28*	.02	.01	.00	.09*	.01	.56	-.01	.05*	.01	.00	-.01
37	.72*	.00	.28	-.36*	.29*	.01	.01	.01	.10*	.00	.56	.00	.06*	.01	.00	.00
38	.90*	.00	.58	-.66*	.26*	.00	.01	.01	.04	.02	.55	-.02	.04	.00	.01	.02
39	.71*	.00	.28	-.48*	.31*	.00	-.01	-.01	.09*	.01	.54	-.04	.06*	.00	-.01	-.01
40	.84*	.01	.28	-.47*	.29*	.04*	.03	.06*	.05	.08*	.51	.01	.03	.04*	.02	.05*
41	.79*	.00	.23	-.47*	.28*	.01	.01	.01	.03	.01	.50	.02	.01	.00	.01	.01
42	.86*	.00	.39	-.61*	.30*	.00	.00	.01	.07*	.01	.50	-.02	.05*	.01	-.02	.00
43	.87*	.00	.42	-.61*	.31*	.00	.00	.01	.07	.01	.50	-.01	.05	.00	-.01	.02

44	.85*	.00	.35	-.61*	.29*	.00	.00	.02	.09*	.02	.49	-.06*	.08*	.01	-.01	.03

Note. R^2 is for step 1, which includes M and M^2. ΔR^2 is the change after adding D1, D2, and D3. *b* is the unstandardized regression coefficient, calculated from the final model. D1, D2, and D3 represent grade 9 vs. grade 11, grade 12, and grade 13, respectively.

* $p < .05$.

Table 29

Incremental R^2s for regression of agreement on complexity and tenure: CT.

	r_{wg}			a_{wg}		
Item	Step 1	Step 2	Step 3	Step 1	Step 2	Step 3
1	.52*	.02	.00	.12*	.07*	.02
2	.42*	.03	.02	.09*	.05	.04
3	.24*	.02	.01	.01	.05	.02
4	.24*	.04	.03	.01	.08*	.05
5	.29*	.01	.00	.02	.03	.01
6	.08*	.03	.02	.02	.02	.03
7	.22*	.03	.00	.03	.01	.00
8	.01	.06	.05	.01	.05	.06
9	.02	.05	.06	.02	.04	.06
10	.11*	.03	.04	.01	.04	.05
11	.02	.04	.09*	.00	.03	.10*
12	.01	.02	.06	.00	.03	.06
13	.28*	.02	.03	.01	.03	.03
14	.10*	.04	.09*	.01	.03	.10*
15	.16*	.02	.04	.01	.05	.05
16	.46*	.01	.04	.02	.05	.06
17	.17*	.02	.06	.00	.02	.08
18	.27*	.05	.05	.04	.03	.08
19	.05	.04	.03	.01	.03	.04

20	.16*	.03	.07	.00	.02	.08
21	.46*	.01	.04	.02	.04	.09*
22	.67*	.01	.01	.06*	.05	.02
23	.46*	.01	.01	.00	.08*	.04
24	.29*	.02	.03	.01	.01	.05
25	.24*	.02	.03	.03	.03	.04
26	.39*	.01	.03	.01	.02	.06
27	.23*	.02	.04	.07*	.00	.05
28	.23*	.02	.08*	.02	.01	.08
29	.30*	.03	.04	.03	.02	.07
30	.12*	.03	.04	.04	.03	.04
31	.64*	.03*	.04*	.23*	.03	.06

Note. Mean importance and its square are entered on Step 1. The main effects for complexity and tenure are entered on Step 2. The complexity × tenure interactions are entered on Step 3.

* $p < .05$.

Table 30

Incremental R^2s for regression of agreement on complexity and tenure: PA.

	r_{wg}			a_{wg}		
Item	Step 1	Step 2	Step 3	Step 1	Step 2	Step 3
1	.67*	.00	.00	.13*	.03	.09*
2	.64*	.00	.01	.04*	.02	.01
3	.43*	.00	.00	.00	.01	.01
4	.31*	.01	.01	.03*	.03	.02
5	.62*	.01	.00	.03	.05*	.01
6	.13*	.01	.01	.01	.04	.01
7	.26*	.01	.02	.03*	.03	.02
8	.43*	.01	.00	.00	.04	.01
9	.40*	.00	.01	.01	.04	.01
10	.35*	.01	.01	.02	.03	.01
11	.10*	.01	.01	.06*	.00	.01
12	.11*	.00	.01	.01	.03	.01
13	.41*	.00	.01	.02	.01	.02
14	.39*	.01	.01	.02	.01	.04
15	.62*	.01	.01	.17*	.01	.03
16	.37*	.01	.01	.00	.02	.01
17	.58*	.00	.01	.01	.04	.03
18	.31*	.00	.00	.03*	.01	.01
19	.36*	.02	.01	.01	.02	.01

20	.49*	.03*	.01	.02	.05*	.01
21	.28*	.06*	.01	.05*	.04*	.01
22	.19*	.02	.01	.09*	.01	.01
23	.13*	.03*	.01	.01	.04	.01
24	.58*	.01	.01	.14*	.03	.01
25	.21*	.01	.01	.01	.04*	.01
26	.05*	.03	.02	.01	.04	.02
27	.09*	.04*	.02	.00	.04*	.02
28	.18*	.01	.02	.04*	.02	.03
29	.09*	.02	.03	.02	.02	.03
30	.28*	.01	.03	.00	.02	.04*
31	.22*	.01	.02	.00	.01	.04*
32	.27*	.00	.03*	.02	.01	.05*
33	.42*	.01	.01	.07*	.02	.01
34	.11*	.01	.03*	.02	.02	.03
35	.03*	.01	.03	.01	.03	.04*
36	.53*	.01	.01	.07*	.01	.01
37	.46*	.01	.00	.05*	.00	.02
38	.73*	.00	.01	.10*	.05*	.02
39	.54*	.01	.00	.04*	.02	.01
40	.61*	.02*	.00	.06*	.05*	.02
41	.62*	.01	.00	.04*	.03	.02
42	.69*	.01	.00	.12*	.03	.02

43	.71*	.01	.01	.13*	.02	.02
44	.72*	.00	.00	.12*	.01	.01

Note. Mean importance and its square are entered on Step 1. The main effects for complexity and tenure are entered on Step 2. The complexity × tenure interactions are entered on Step 3.

* $p < .05$.

Table 31

Tenure-complexity cell means for items with significant interactions: CT.

		Grade										
		4		5		6		7		8		
Item	Tenure	M	SD	M	SD	M	SD	M	SD	M	SD	*F*
r_{wg}												
11	Low	.11	.39	.29	.25	.24	.25	.22	.31	.52	.32	4.56*
	Medium	.21	.22	.26	.14	.19	.16	.28	.21	.16	.21	.88
	High	.35	.15	.29	.10	.27	.10	.28	.10	.29	.12	.28
	F	5.67*		.12		.66		.90		4.43*		
14	Low	.38	.29	.27	.22	.15	.21	.10	.31	.45	.26	6.26*
	Medium	.25	.18	.35	.10	.26	.17	.27	.18	.20	.22	.96
	High	.39	.18	.35	.09	.31	.07	.25	.10	.23	.12	.93
	F	2.73		1.06		2.28		3.58*		2.91		
28	Low	.34	.24	.41	.19	.40	.14	.28	.22	.66	.20	6.38*
	Medium	.39	.17	.38	.12	.42	.12	.47	.12	.50	.25	.32
	High	.43	.14	.40	.10	.47	.09	.49	.11	.53	.12	.38
	F	.73		.13		1.46		6.00*		2.65		
31	Low	.62	.36	.56	.24	.55	.20	.55	.35	.73	.12	8.33*
	Medium	.60	.16	.60	.14	.62	.14	.68	.12	.67	.16	.20
	High	.63	.14	.64	.10	.64	.10	.67	.10	.65	.13	.81
	F	1.96		.75		3.35		9.63*		2.05		
a_{wg}												

11	Low	.54	.22	.66	.09	.64	.12	.62	.14	.76	.15	4.82*
	Medium	.61	.09	.63	.07	.60	.07	.65	.09	.59	.11	.72
	High	.68	.08	.64	.05	.63	.06	.64	.04	.64	.06	.54
	F	5.26*		.39		.83		.18		5.09*		
14	Low	.66	.15	.64	.11	.60	.10	.57	.14	.75	.12	5.09*
	Medium	.61	.08	.66	.05	.63	.08	.64	.09	.61	.11	.68
	High	.68	.09	.66	.04	.64	.04	.62	.04	.62	.06	1.25
	F	2.90		.22		.57		1.49		4.43*		
21	Low	.75	.10	.73	.09	.73	.10	.75	.05	.88	.09	3.04*
	Medium	.73	.06	.72	.06	.72	.03	.74	.06	.72	.06	.29
	High	.73	.05	.73	.04	.73	.04	.73	.03	.72	.05	.12
	F	.95		.06		.06		.13		7.15*		

Note. *F* values are for simple effects of tenure within grade and grade within tenure.

* *p* < .05

Table 32

Tenure-complexity cell means for items with significant interactions: PA.

		Grade								
		9		11		12		13		
Item	Tenure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*
r_{wg}										
32	Low	.59	.22	.52	.26	.40	.31	.26	.39	5.62*
	Medium	.45	.20	.50	.14	.47	.22	.47	.26	.68
	High	.45	.18	.47	.13	.45	.12	.45	.14	.44
	F	3.05		.10		1.74		3.27*		
34	Low	.61	.18	.58	.17	.52	.24	.46	.35	2.83*
	Medium	.57	.13	.59	.09	.54	.17	.65	.16	2.09
	High	.56	.11	.56	.08	.57	.05	.58	.06	.70
	F	.82		.67		1.10		6.16*		
a_{wg}										
1	Low	.70	.12	.70	.17	.80	.12	.28	.11	13.20*
	Medium	.69	.09	.70	.08	.67	.12	.73	.19	1.27
	High	.67	.10	.68	.07	.68	.05	.69	.08	.35
	F	.62		.86		5.44*		2.67		
30	Low	.73	.11	.71	.12	.62	.13	.74	.15	6.60*
	Medium	.68	.08	.69	.07	.70	.09	.71	.11	.92
	High	.67	.08	.69	.06	.68	.05	.69	.06	.44
	F	3.68*		.39		2.93		.22		

31	Low	.75	.11	.77	.11	.69	.11	.69	.17	5.43*
	Medium	.74	.07	.74	.05	.74	.07	.75	.08	.21
	High	.72	.07	.73	.04	.74	.03	.74	.05	.50
	F	1.35		2.23		1.31		2.21		
32	Low	.74	.12	.73	.14	.64	.16	.64	.16	7.07*
	Medium	.66	.09	.69	.06	.69	.10	.70	.12	1.46
	High	.67	.08	.67	.06	.67	.06	.68	.07	.32
	F	6.92*		1.69		.75		1.81		
35	Low	.70	.10	.73	.11	.69	.15	.65	.20	4.67*
	Medium	.67	.09	.68	.08	.68	.09	.73	.13	3.57*
	High	.67	.08	.67	.05	.67	.06	.68	.07	.21
	F	3.64		4.59*		1.42		5.92*		

Note. *F* values are for simple effects of tenure within grade and grade within tenure.

* *p* < .05

Table 33

Regression of agreement on KSAO abstractness.

	Step ΔR^2			b				
	1	2	3	M	M^2	A	D	A×D
CT								
r_{wg}	.71*	.01*	.00*	.30*	.14*	.03*	-.01	.01*
a_{wg}	.20*	.04*	.01*	.04*	-.01	.01*	-.00 *	.01*
PA								
r_{wg}	.73*	.00	.02*	.21*	.18*	.02*	-.00	.05*
a_{wg}	.43*	.00	.02*	.06*	-.01*	.00*	.00	.02*

Note. Based only on items with acceptable agreement of abstractness. M = centered mean importance. A = abstractness. D = Multidimensionality. Step 1 enters M and M^2. Step 2 enters A and D. Step 3 enters A×D. All *b*s are for the final model.

* $p < .05$.

Table 34

Consequences of removing disagreement.

	r_{wg}			a_{wg}					
SD Group	*M*	*SD*	% Good	*M*	*SD*	% Good	% OOR	% N	% Crit.
CT									
All	.48	.16	6.83	.68	.06	1.02	0.00	—	—
2.00	.56	.18	27.21	.72	.06	7.22	0.15	3.11	6.53
1.75	.60	.16	31.69	.75	.06	13.96	0.23	5.62	9.75
1.50	.67	.14	44.97	.78	.06	35.20	0.27	10.22	10.36
1.25	.77	.11	69.15	.83	.07	66.35	0.30	19.87	7.51
1.00	.87	.08	95.29	.91	.06	96.42	0.49	36.08	12.41
0.75	.94	.06	99.92	.97	.04	99.54	0.57	53.65	11.39
0.50	.99	.04	100.00	.99	.02	100.00	6.89	67.31	9.56
0.25	1.00	.00	100.00	1.00	.00	100.00	79.99	81.57	30.06
PA									
All	.45	.23	12.25	.63	.10	0.70	0.26	—	—
2.00	.56	.25	35.81	.70	.09	6.65	0.41	3.77	2.45
1.75	.61	.24	40.99	.72	.09	12.51	0.41	6.07	4.86
1.50	.68	.19	49.83	.76	.08	26.38	0.39	10.12	6.34
1.25	.76	.14	65.84	.80	.08	50.02	0.38	17.47	4.48
1.00	.86	.10	87.86	.87	.09	78.46	0.44	30.04	6.87
0.75	.94	.06	99.52	.95	.07	91.88	0.57	49.42	6.90
0.50	.99	.04	100.00	.99	.02	99.93	11.09	68.81	9.83

0.25	1.00	.00	100.00	1.00	.00	100.00	90.46	84.41	26.58

Note. For SD Group, All = the entire dataset without deletion; other rows exclude cases whose rating is more than the indicated standard deviations from the mean. % Good = the percentage of values that are greater than .70 (r_{wg}) or .80 (a_{wg}). % OOR = the percentage of a_{wg} values that can not be interpreted. % N = the average percentage of cases removed. % Crit. = the percentage of items whose criticality ($M \geq 3.5$) changes after removing cases.

Figure 1

Sources of disagreement.

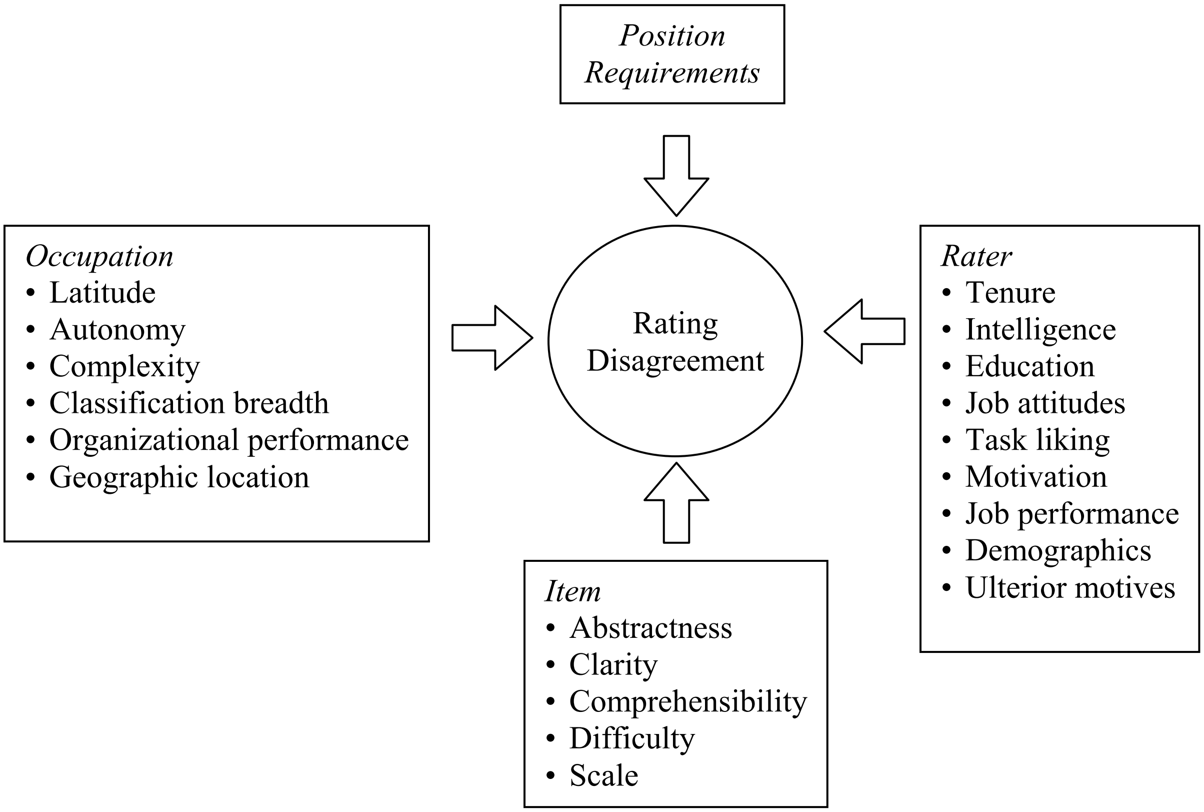
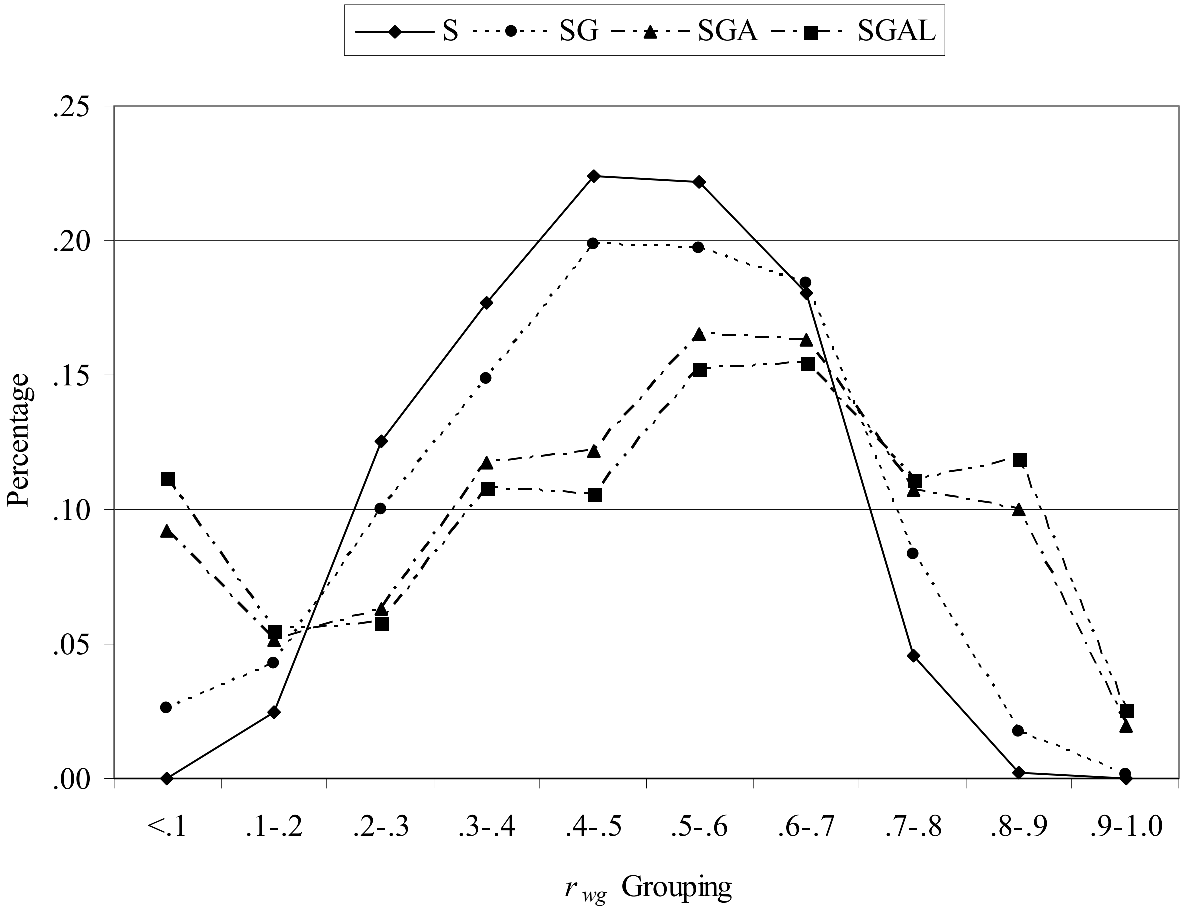


Figure 2

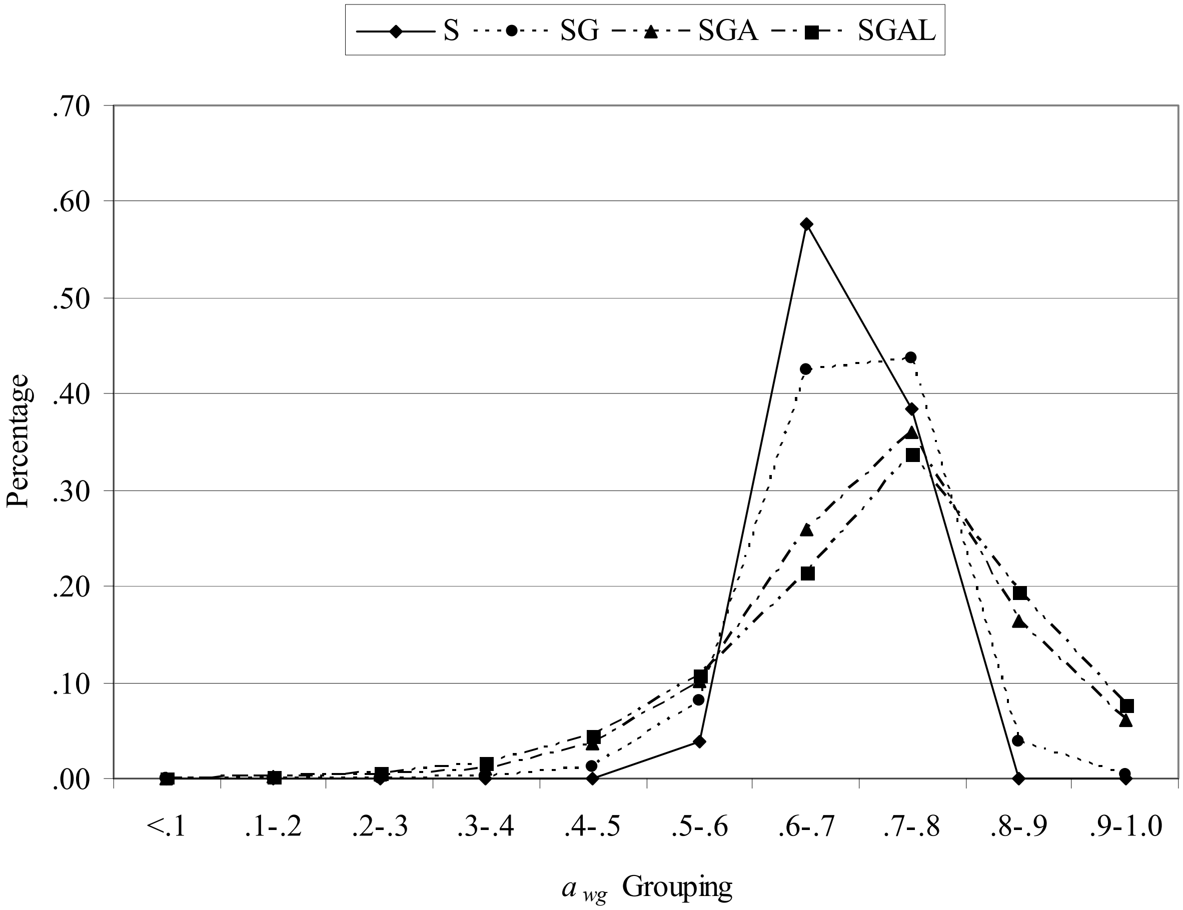
Distribution of r_{wg} for CT Occupations by level of aggregation.



Note. S = series; SG = series-grade; SGA = series-grade-agency; SGAL = series-grade-agency-location.

Figure 3

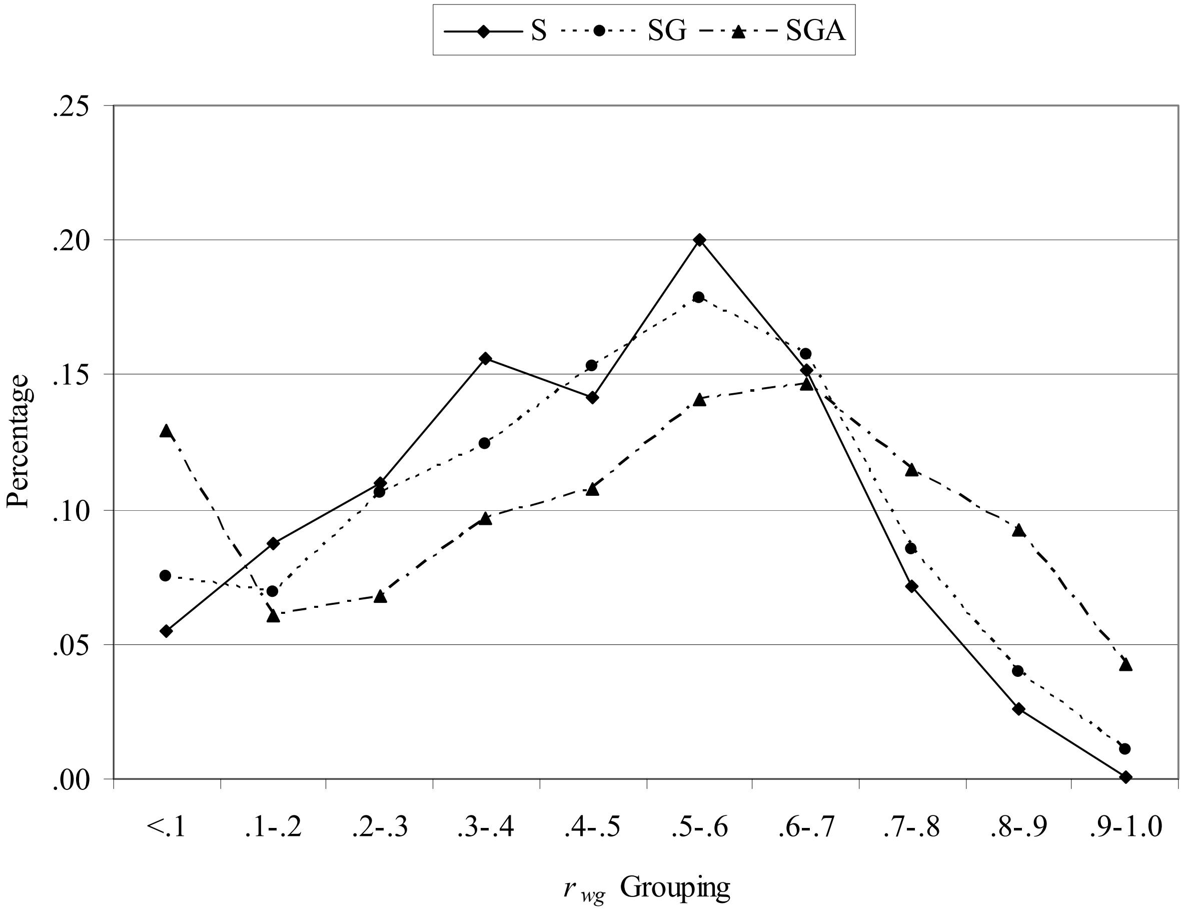
Distribution of a_{wg} for CT Occupations by level of aggregation.



Note. S = series; SG = series-grade; SGA = series-grade-agency; SGAL = series-grade-agency-location.

Figure 4

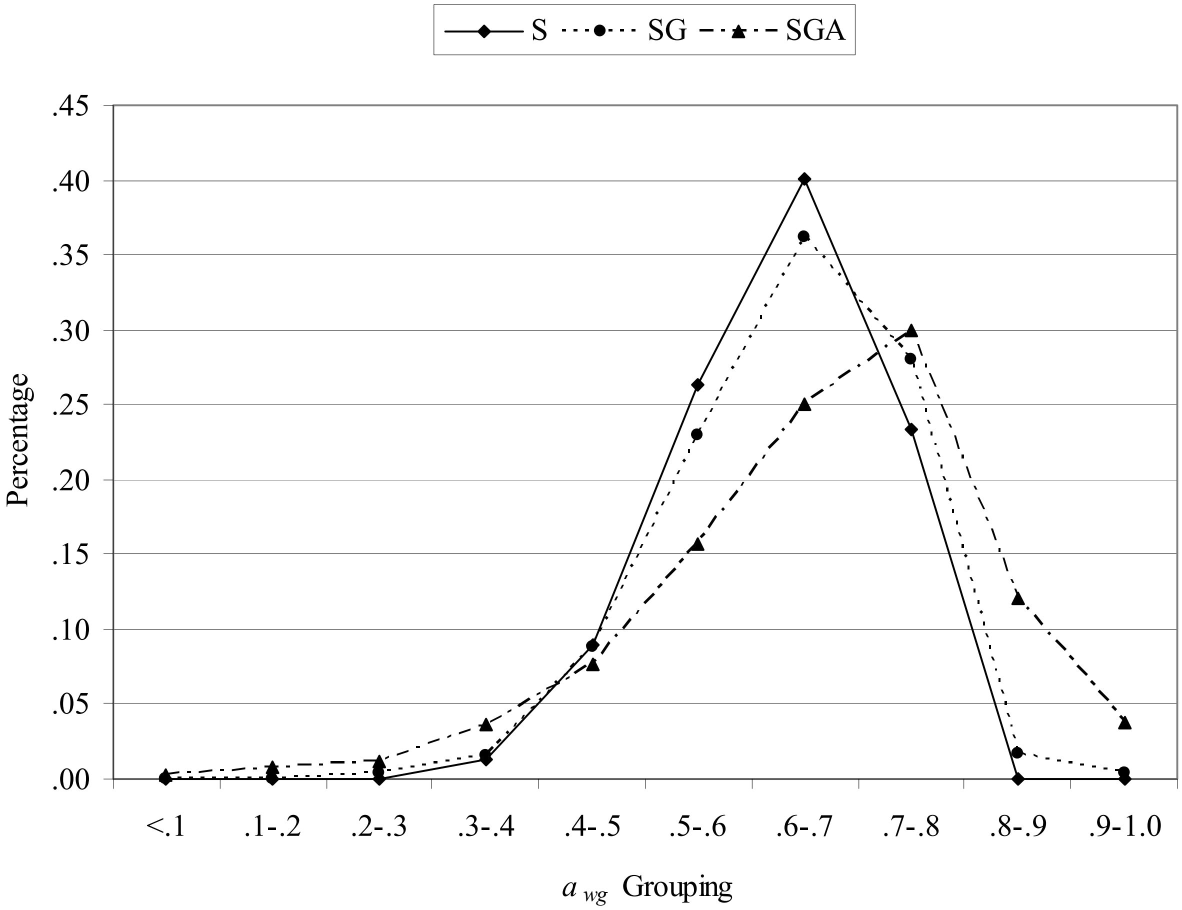
Distribution of r_{wg} for PA Occupations by level of aggregation.



Note. S = series; SG = series-grade; SGA = series-grade-agency.

Figure 5

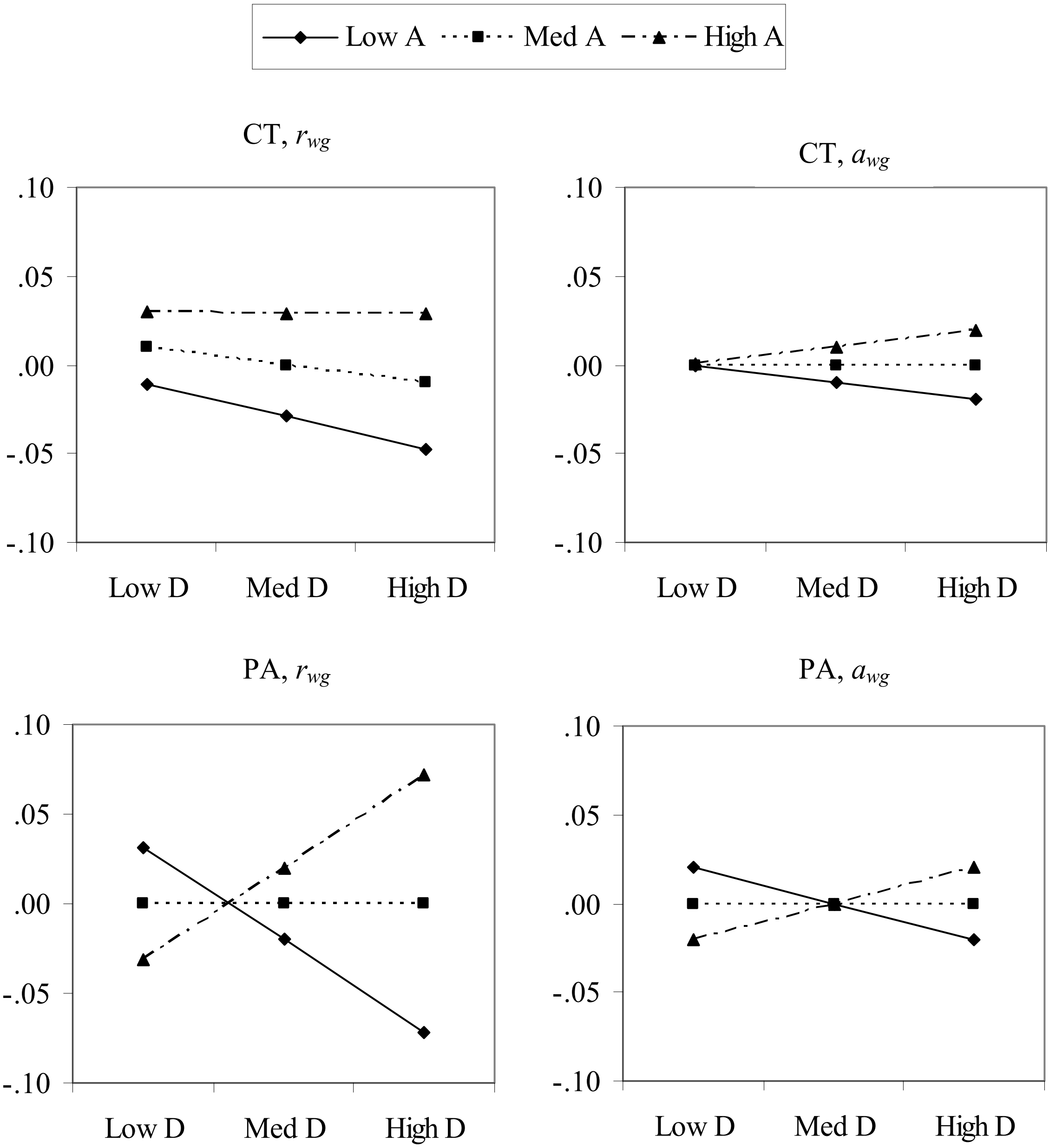
Distribution of a_{wg} for PA Occupations by level of aggregation.



Note. S = series; SG = series-grade; SGA = series-grade-agency.

Figure 6

Interaction of Construct Abstractness and Definition Multidimensionality.



Note. A = Construct abstractness. D = Definition multidimensionality. Low = 1 standard deviation below the mean. Med = the mean. High = 1 standard deviation above the mean. Based only on items with acceptable agreement of abstractness, assuming a KSAO of average importance.

Steven R. Burnkrant

Education

- *Ph.D., Industrial and Organizational Psychology, Virginia Tech, 2003.*
- *M.S., Industrial and Organizational Psychology, Virginia Tech, 1999.*
- *B.A., Psychology, The Colorado College, 1996.*

Experience

- *Personnel Research Psychologist, U.S. Office of Personnel Management, 2000 – present.*
  Customer Satisfaction: Project director. Developed surveys. Conducted SME panels and focus groups. Briefed agency division heads on results. Wrote a program to automatically score and produce reports.
  Leadership 360˚ Feedback: Project director. Provided group and individual feedback. Headed team that developed a new instrument. Conducted preliminary validation. Wrote program to automatically score and produce reports.
  Program Evaluation: Facilitated SME focus groups on the effectiveness of a large Federal demonstration project, analyzed survey and personnel data, wrote technical report.
  Job analysis: Analyzed data and wrote report on subset of Federal clerical and technical, professional and administrative, and managerial occupations. Member of team redesigning the MOSAIC methodology.
  Classification: Developed a simplified factor evaluation model using multivariate statistics and confirmatory factor analysis, briefed Federal classifiers, facilitated panels of SMEs in validation test, analyzed data, wrote technical report.
  Selection: Facilitated SME panels to identify core competencies for assessment, analyzed ratings, wrote technical report. Oversaw development and implementation of a web-based assessment.
- *Instructor, Department of Psychology, Virginia Tech, 1999 – 2000.*
  Taught sophomore-level Social Psychology and senior-level Industrial and Organizational Psychology.
- *Research Assistant, Educational Technologies, Virginia Tech, 1997-1999.*
  Developed and administered surveys designed to assess the impact of technology on learning. Conducted quantitative and qualitative data analyses. Wrote technical reports.
- *Research Assistant, Department of Engineering, Virginia Tech, 1996 – 1997.*
  Conducted archrival (N > 100,000) data analysis and wrote a report on the performance of women and minorities in undergraduate engineering programs.

Computer Skills

- *Statistical: SPSS, SAS, AMOS, LISREL, BILOG, WordStat.*
- *Other: Word, Excel, PowerPoint, Access, Visual Basic, HTML, JavaScript.*

Awards

- *Sustained Superior Performance Award (1/02; 1/03); Spot Cash Award (2/01, 3/01, 9/01, 10/02, 1/03); Pendleton Award (6/02); Gold Star Award (6/02); Special Act Award (8/02, 8/02, 7/03); Director's Award for Excellence (8/02).*

Professional Affiliations

- *Society for Industrial and Organizational Psychology.*
- *Academy of Management.*
- *American Psychological Association.*

Research

- Burnkrant, S. R. (2003). *Interrater agreement of incumbent job specification importance ratings: Rater, occupation, and item effects.* Unpublished doctoral dissertation, Virginia Tech.
- Ford, J., & Burnkrant, S. R. (2002, October). *Applications of automated text analysis technology*. Symposium presented at The 30th International Congress on Assessment Center Methods, Pittsburg, PA.
- Burnkrant, S. R. (2002, April). Linking merit pay to motivation, organizational commitment, turnover intentions, and performance. In R. L. Heneman (Chair), *Merit Pay Revisited: Is it effective? Is it fair?* Symposium conducted at the 17th annual conference of the Society for Industrial and Organizational Psychology, Toronto, Canada.
- Burnkrant, S. R., & Thibodeaux, H., Niles-Jolly, K., & Paskey, E. (2002). *Development of the Administrative Law Judge competency-based job profile*. Technical Report, Personnel Resources and Development Center, U.S. OPM.
- Chmielewski, M. Medley-Proctor, K., Burnkrant, S. R., Schay, B., & Buckley, T. (2001). *Validation of fewer than nine factors in a leveling tool.* Technical Report, Personnel Resources and Development Center, U.S. OPM.
- Shoun, S., Burnkrant, S. R., & Pick, S. (2001). *Occupational Analysis of selected Library of Congress Occupations: An applicant of the Multipurpose Occupational Systems Analysis Inventory—Closed-Ended (MOSAIC).* Technical Report, Personnel Resources and Development Center, U.S. OPM.
- Beach, M. A., Buckley, T., Burnkrant, S. R., et al. (2001). *2000 implementation report: DoD S&T Reinvention Laboratory Demonstration Program.* Technical Report, Personnel Resources and Development Center, U.S. OPM.
- Burnkrant, S. R. (2001, April). *Effects of Competition on faking job-specific profiles: Job-desirability or social desirability?* Paper presented at the 16 Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.
- Burnkrant, S. R., & Taylor, C. D. (2001, April). *Equivalence of traditional and internet-based data collection: Three multigroup analyses*. Paper presented at the 16 Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.
- Burnkrant, S. R., & Foti, R. J. (2000). *Antecedents, correlates, and cross-domain consistency of learning, prove, and avoid goals.* Unpublished Masters Thesis, Virginia Tech.
- Burnkrant, S. R., & Harvey, R. J. (2000, April). *Establishing base rates for the Z_3 and F_2 inappropriateness indices for use with the Meyers-Briggs Type Indicator.* Paper presented at the 15 Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.
- Burnkrant, S. R. (1999, April). *Relationships among commitment foci: Consequences for organizational behavior and attitudes.* Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.
- Burnkrant, S. R., & Hauenstein, N. M. A. (1999). *Modeling a nonrecursive relation between procedural and distributive justice: A reanalysis of Tyler (1994).* Unpublished manuscript.
- Taylor, C. D., & Burnkrant, S. R. (1999). *Spring 1999 online courses: Assessment report to the Institute for Distance and Distributive Education*. Technical report, Virginia Polytechnic Institute and State University.
- Taylor, C. D., & Burnkrant, S. R. (1999). *Summer 1999 online courses: Assessment report to the Institute for Distance and Distributive Education.* Technical report, Virginia Polytechnic Institute and State University.
- Taylor, C. D., & Burnkrant, S. R. (1998). *On-line summer school 1998: Assessment report.* Technical report, Virginia Polytechnic Institute and State University.
- Taylor, C. D., & Burnkrant, S. R. (1998). *Maymester 1998: Assessment report.* Technical report, Virginia Polytechnic Institute and State University.
- Taylor, C. D., & Burnkrant, S. R. (1998). *Comparative assessment of a traditional class, a studio class, and a distant class.* Technical report, Virginia Polytechnic Institute and State University.