TRANSIENT ERROR AND COEFFICIENT ALPHA: A CALL FOR CAUTIOUS PRACTICE WHEN APPLYING AND INTERPRETING ALPHA IN PERSONNEL SELECTION SETTINGS

Except where reference is made to the work of others, the work described in this dissertation is my own or was done in collaboration with my advisory committee. This does not include proprietary or classified information.

_____
Christopher S. Winkelspecht

Certificate of Approval:

_____	_____
Philip Lewis Professor Psychology	Adrian Thomas, Chair Associate Professor Psychology
_____	_____
John G. Veres, III Professor Psychology	Frank Weathers Associate Professor Psychology

_____
Joe F. Pittman
Interim Dean
Graduate School

TRANSIENT ERROR AND COEFFICIENT ALPHA: A CALL FOR CAUTIOUS PRACTICE WHEN APPLYING AND INTERPRETING ALPA IN PERSONNEL SELECTION SETTINGS

Christopher S. Winkelspecht

A Dissertation

Submitted to

the Graduate Faculty of

Auburn University

in Partial Fulfillment of the

Requirements for the

Degree of

Doctor of Philosophy

Auburn, Alabama
December 15, 2006

TRANSIENT ERROR AND COEFFICIENT ALPHA: A CALL FOR CAUTIOUS PRACTICE WHEN APPLYING AND INTERPRETING ALPHA IN PERSONNEL SELECTION SETTINGS

Christopher S. Winkelspecht

_____
Signature of Author

_____
Date of Graduation

DISSERTATION ABSTRACT

TRANSIENT ERROR AND COEFFICIENT ALPHA: A CALL FOR CAUTIOUS PRACTICE WHEN APPLYING AND INTERPRETING ALPHA IN PERSONNEL SELECTION SETTINGS

Christopher S. Winkelspecht

Doctor of Philosophy, December 15, 2006
(M.S., Auburn University, 2003)
(B.A., Temple University, 1999)

91 Typed Pages

Directed by Adrian Thomas

Reliability is an integral component in determining the worth of results from any measure. There are a number of estimates used to represent reliability, which vary in terms of the sources of error addressed, underlying assumptions about the data, statistical theory, and formulae applied; but in the areas of personnel selection research and practice coefficient alpha (also known as Cronbach's alpha and simply referred to as alpha below) is by far the most widely reported. Alpha's popularity is mostly due to two commonly accepted properties of the statistic. First, it is a measure of internal consistency so, unlike test-retest or inter-rater estimates of reliability, data used to generate the coefficient can be gathered from a single test administration. Second, alpha is generally considered a conservative statistic, or more specifically the coefficient is thought to estimate reliability's lower boundary. While the convenience of calculating alpha is inarguable,

the assumption that it is an underestimate of reliability is not always warranted. It has recently been demonstrated that transient errors, a source of variability not often assessed, can actually inflate the alpha coefficient and cause reliability to be overestimated. The current study investigates the effect of transient error and echoes the call to present additional diagnostic information that has recently been introduced to the professional literature, such as Alpha's Standard Error (ASE; Duhacheck & Iacobucci, 2004). The benefits of calculating confidence intervals surrounding the alpha coefficient and substituting the upper and lower boundaries in place of the point estimate when performing a variety of calculations used in personnel selection practice and research are demonstrated. The present research calls for greater caution interpreting and applying reliability estimates in this high stakes setting.

ACKNOWLEDGEMENTS

The author would like to share his gratitude for those who helped make this work possible. Many thanks are due to Dr. Adrian Thomas for his countless time and contributions as chair of this dissertation committee. Thank you Dr. John G. Veres, III for all of the knowledge and experiences you have shared. Dr. Frank Weathers, thank you for your continued support and service on both my thesis and dissertation committees. Thank you Dr. Allison Jones-Farmer for the expertise and feedback you have provided. Dr. Philip Lewis, thank you for all of the support you have given me and the entire Industrial/Organizational psychology program at Auburn. Both this research and I have greatly benefited from the support and assistance this committee has provided. I am greatly appreciative.

Special thanks to my mother, Susan, for her lifelong love and devotion, without which I would have never been able to pursue this degree. Likewise, thanks to my sister, Kathleen, who always cheers me on and up. Thanks to my wonderful wife, Cami, for the love, patience, and encouragement she constantly provides. Finally, I would also like to extend my gratitude to all my family and friends who have provided boundless support throughout my tenure as a graduate student. So many people have helped me in so many ways. I am neither able to thank you all here nor thank you all enough, but I deeply appreciate your generous support.

Style manual or journal used:

Publication Manual of the American Psychological Association, 5th edition

Computer software used:

Microsoft Word 2003, Statistical Package for the Social Sciences (SPSS), 11.5, Microsoft Excel 2003, and PROC MULTEVENT.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Reliability is an integral component in determining the worth of results from any measure. There are a number of estimates used to represent reliability, which vary in terms of the sources of error addressed, underlying assumptions about the data, statistical theory, and formulae applied; but in the areas of personnel selection research and practice coefficient alpha (also known as Cronbach's alpha and simply referred to as alpha below) is by far the most widely reported. Alpha's popularity is mostly due to two commonly accepted properties of the statistic. First, it is a measure of internal consistency so, unlike test-retest or inter-rater estimates of reliability, data used to generate the coefficient can be gathered from a single test administration. Second, alpha is generally considered a conservative statistic, or more specifically the coefficient is thought to estimate reliability's lower boundary. While the convenience of calculating alpha is inarguable, the assumption that it is an underestimate of reliability is not always warranted. It has recently been demonstrated that transient errors, a source of variability not often assessed, can actually inflate the alpha coefficient and cause reliability to be overestimated. The current study investigates the impact of transient error and echoes the call to present additional diagnostic information that has recently been introduced to the professional literature, such as Alpha's Standard Error (ASE; Duhacheck & Iacobucci, 2004). The benefits of calculating confidence intervals surrounding the alpha coefficient and substituting the upper and lower boundaries in place of the point estimate when

performing a variety of calculations used in personnel selection practice and research are demonstrated. The present research calls for greater caution interpreting and applying reliability estimates in this high stakes setting.

Personnel Selection and Test Reliability

Personnel selection is the process through which individuals are identified and hired to fill vacancies within an organization. In order to comply with federal regulations, organizations create selection systems in accordance with various laws (e.g., Title I of the Civil Rights Act of 1991; The Americans with Disabilities Act, 1990) and following professional guidelines (e.g., The Uniform Guidelines on Employee Selection, 1978; The Principles for the Validation and Use of Personnel Selection Procedures, 2003). The ultimate goal of a personnel selection system is to hire individuals who will provide a return on the organization's investments in them. A considerable amount of time, money, and other resources are spent recruiting, selecting, training, and compensating individuals to perform services for the organization. An organization's selection procedures help identify the people who have the knowledge, skills, and abilities (KSAs) important to successfully perform the duties of the positions being filled (either immediately or after training). There are numerous selection devices that can be used to assess applicants' KSAs but whatever the technique (e.g., written exam, structured oral interview, role play) the substance of the assessment (i.e., the content and results) must be valid and reliable.

A valid assessment measures what it is intended to measure. If an interview is designed to gauge applicants' management capability the content should focus on general management issues and not involve other areas of knowledge, such as finance, or other

abilities, such as mathematical aptitude. The results should also correlate with job-related variables such as training proficiency and/or job performance. Furthermore, it is critically important that results be reliable, such that if an identical assessment were to be administered the applicants' performances would be replicated. Unfortunately, no test is perfect; irrelevant factors are certain to influence results, thus, assessments can neither be perfectly valid nor perfectly reliable. Irrelevant factors stem from a wide variety of sources ranging from internal elements of the test (such as poorly worded questions) to external elements of the environment (such as the degree to which others are observing the test-taker). Any source of influence unrelated to the area of knowledge, skill, or ability being assessed is considered measurement error. Measurement error confounds the interpretation of test results and can significantly decrease an organization's ability to choose the most highly qualified individual(s) from among less desirable applicants. The ability to detect the presence of measurement error stems from reliability research within the area of psychometric and statistical theory.

Reliability Estimates

Statistical formulae used to assess the reliability of a measure have been generated on the basis of several theoretical models, but the most influential has been the Classical True Score Model also more simply known as Classical Test Theory (CTT). CTT's origins began during Spearmen's work with the correlation coefficient in the early 1900's. In a series of publications from 1904-1913 Spearmen presented logical and mathematical proofs that scores from any test are inaccurate measures, to a certain degree, of whatever trait is being assessed (Crocker & Algina, 1987). The basis of the CTT model rests on Spearmen's assertion that any test score can, and should, be

considered the composite of two hypothetical elements: a true score and some quantity of random error.

$$X_o = T_x + E_x \quad \text{(Equation 1)}$$

Based on this model (where X_o represents the observed score, T_x is the true score, and E_x represents measurement error), reliability coefficients are formulated to represent the extent to which examinees' scores on a test covary with the "true" extent to which they possess the knowledge, skills, or abilities that are related to that being tested. The correlation between observed and true scores is estimated by taking the square root of the reliability estimate (Equation 2).

$$\sqrt{r_{xx}} = r_{XT_X} \quad \text{(Equation 2)}$$

For example, if .85 is found to be the reliability coefficient for a measure of conscientiousness this suggests that individuals' observed scores are linearly correlated with their true scores at .92 ($\sqrt{.85} = .92$).

There are other theories of reliability, such as Generalizability Theory and Domain Sampling, which propose models with slightly different emphases, usually both theoretically and mathematically. The convergences and contrasts among the different theories and their applications is beyond the scope of the current work but it should be kept in mind that the discussion below would be presented somewhat differently if a model other than CTT served as the theoretical basis. Even within CTT there are numerous ways to estimate the reliability of scores (e.g., internal consistency, test-retest, form equivalence; Thompson, 2003). The present work will specifically focus on variants of the parallel test approach under CTT (namely, the internal consistency method). Creating parallel tests is a general strategy for assessing the stability of an

individual attribute. The method is to develop two forms of a test that produce consistent (with respect to rank-order) true scores for people on both the first and second version (parallel tests). After individuals complete both forms their scores can be compared and any differences can be attributed to measurement error. Great rank-order differences between scores on the two tests should raise concerns about measurement error, while consistent rank-order should foster confidence in the stability of the results. Strictly parallel tests are extremely difficult to create and so this approach mostly serves as a conceptual foundation for more practical methods in research and practice. These methods include the test-retest correlation and various derivations of the internal consistency method. However, applied personnel practitioners rarely have the data to investigate test-retest correlations, so the focus of the present work will center on the internal consistency method.

The Internal Consistency Method and Coefficient Alpha

The internal consistency method offers a fairly simple answer to the question "what is reliability" if each item on a test is considered a distinct behavioral observation. The more observations one takes and the greater the consistency among those observations, the higher the reliability estimate (Murphy & Davidshofer, 2001). Internal consistency estimates are dependent upon the number of observations reported (i.e., items) and the extent to which those observations covary from instance to instance (i.e., the intercorrelations among those items). Coefficient alpha is the most commonly generated statistic of the internal consistency method.

Cronbach (1951) originally introduced coefficient alpha as an extension to one of Kuder and Richardson's (1937) reliability estimates, known as Kuder-Richardson #20 (or

KR20). Kuder and Richardson presented several formulae designed to summarize the reliability of a test with more than one item, which are all dichotomously scored (i.e., 0 = incorrect and 1 = correct; Knapp, 1991). The best known and most commonly referenced is the 20th equation presented in that work. The KR20 formula is:

$$KR20 = k/(k-1)[1 - \sum p_i(1 - p_i / s_T{}^2)] \quad \text{(Equation 3)}$$

Where k is the number of items, p_i is the proportion of people with a score of 1 on the k[th] item, and $s_T{}^2$ is the variance of the scores on the total test.

After the introduction of the KR20 statistic, Hoyt (1941) demonstrated that the same general calculation could be produced through a repeated-measures analysis of variance approach to the subject-by-item data matrix. Several years later Cronbach (1951) extended the work to include tests with more than two choices for each item. Both KR20 and alpha are linked to the parallel test approach through the split-half method, where a single test is divided into parts and scores on one part are correlated with the scores on the other (Murphy & Davidshofer, 2001). Specifically, alpha represents the mean correlation that would result from every possible combination of split tests.

The general formula for Cronbach's coefficient alpha is:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_T^2}\right] \quad \text{(Equation 4)}$$

Where k is the number of items in the test, σ_i^2 is the variance of the i[th] item, and $\sigma_T{}^2$ is the total test variance.

There are a variety of commonly accepted interpretations of what alpha measures. For example it has been stated (Cortina, 1993) that alpha is: 1) the mean of all split-half reliabilities (Cronbach, 1951); 2) the lower bound of reliability of a test (Kristoff, 1974; Novick & Lewis, 1967); 3) a measure of first-factor saturation, or unidimensionality (Crano & Brewer, 1973; Hattie, 1985); 4) equal to reliability in conditions of essential tau-equivalance (τ-equivalence); and 5) a more general version of the KR20 coefficient (Cronbach, 1951; Fiske, 1966, Hakstian & Whalen, 1976).

Crocker and Algina (1986) offer the following interpretation (p. 120):

> "When a composite test is made up of nonparallel subtests, we can estimate the lower bound of its coefficient of precision by using coefficient alpha. This computation requires that we know the number of subtests, the variance of the composite scores, and the sum of all the subtest covariances. The usefulness of this relationship will be more apparent if we recall that any test may be regarded as a composite and each item as a subtest. Thus, coefficient alpha provides a convenient way to estimate the lower bound of the coefficient of precision for a test by using item response data obtained from a single administration of that test."

Assumptions of Coefficient Alpha

In order for alpha to be interpreted as an accurate reflection of reliability two assumptions must hold true. First, while the items need not be perfectly parallel they must be essentially τ-equivalent[1]. It has long been known that the assumption of τ-equivalence is routinely violated and mathematically demonstrated that when this occurs alpha will produce a lower bound estimate of reliability (Novick & Lewis, 1967). However, it has been noted that a basis for the calculations that led to this conclusion is that the second assumption must hold true (Zimmerman, Zumbo, & Lalonde, 1993); error associated with individual items must not be correlated with the errors of other items.

[1] Items are defined to be essentially τ-equivalent when true score counterparts differ only by an additive constant.

Zimmerman et al. (1993) investigated the effects of the two violations separately and found that violation of the τ-equivalence assumption produces a deflated reliability coefficient while violation of uncorrelated errors produces an inflated estimate. However, the result of each violation in isolation does not indicate how simultaneous violations of both assumptions will effect the reliability estimate. The possibility existed that one violation trumped the other or that violations of both at the same time creates a wash effect where neither exert enough influence to inflate or deflate the estimate. It was not until Komaroff (1997) investigated the effects of simultaneous violations of essential τ-equivalence and uncorrelated error on coefficient alpha that a solid understanding of the impact was made known and the alpha coefficient could be fully interpreted.

To determine the interactive effect of simultaneous violations of both essential τ-equivalence and uncorrelated error Komaroff (1997) varied true score correlations among items on a hypothetical assessment, error score correlations, and the number of items with correlated error. Results demonstrated that correlated error attenuates the degree to which alpha underestimates $p_{xx'}$,[2] under violations of essential τ-equivalence, and when this effect is most pronounced alpha can overestimate $p_{xx'}$. Komaroff (1997) demonstrated that under violations of these dual assumptions alpha remains likely to be an overestimate of reliability's lower boundary. To demonstrate, take the basic classical test theory linear model for two individual items:

$$X_1 = T_1 + E_1;\ X_2 = T_2 + E_2 \quad \text{(Equation 5)}$$

[2] $p_{xx'} = \sigma^2(T)/\sigma^2(X)$, where $\sigma^2(T)$ is the true composite variance and $\sigma^2(X)$ is the observed composite variance.

Where X = observed score, T = true score and E = error score and the covariance between the two items is:

$$COV(X_1,X_2) = \sigma_{X_1}\,\sigma_{X_2}\,r_{X_1X_2} =$$

$$COV(T_1,T_2) + COV(E_1,E_2) + COV(T_1E_1) + COV(T_2E_2) \qquad \text{(Equation 6)}$$

$COV(T_i, E_i) = 0$, because true score variance cannot be associated with error variance by definition, so the last two terms drop out. $COV(E_1,E_2) = 0$ is an assumption and can be violated by data (Komaroff, 1997b). If $COV(E_1,E_2) > 0$, the sum of observed (X) item covariances will be inflated and coefficient alpha will be an overestimate of reliability. Returning to the formula for coefficient alpha (Equation 4), recall σ_T^2 is the is the sum of all the test items' variances and covariances. If the sum of the covariances is inflated, σ_T^2 will be inflated, which in turn will decrease the value of $\frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_T^2}$, increase the value of $\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_T^2}\right]$, and, to whit, cause alpha $\frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_T^2}\right]$ to be an inflated estimate of internal consistency.

Like most assumptions used as a basis for operation, the accuracy of $COV(E_1,E_2) = 0$ has not been greatly scrutinized. A potential reason for this oversight is evident returning to the Crocker and Algina reference (above). If alpha is indeed used as a calculation of the consistency among test scores (i.e., alpha reflects reliability among

separate tests, not items[3]), it is reasonable to assume that whatever irrelevant factors contributed to performance on one test are not likely to be related to irrelevant factors that influenced performance on other tests. For example, if construction work is being performed outside of a classroom on a warm day and the air conditioning is out of order forcing the teacher to leave the window open throughout the testing period, this may negatively affect an individual's performance on that test but the situation is not likely to reappear during other administrations. In this instance the error variance associated with test conditions would be relatively uncorrelated. In contrast, if alpha is used as a calculation of the consistency among items on that individual test, the noise could affect the student's answers to any or all of the test questions, as the noise would be present from one to the next. In this instance, hypothetically different "administrations" would have correlated sources of error variance. This may be a poor example to present to testing specialists who meticulously control locations to eliminate such "noise." Yet even experts can only control environmental factors. The influence of internal factors, such as an individual's psychological state, cannot be so easily manipulated as the window. These transient errors, as they have been termed (Becker, 2000), which consist of fluctuations in test-takers' moods/affect/states, while random from one test occasion to another, might saturate many (if not all) of the items an individual answers throughout a single test administration.

Transient Error

Transient errors are response variations that are due to random changes in test-takers' psychological states across time. As the changes are not related to the construct(s)

[3] Alpha can represent the internal consistency of data at the test or item level.

being assessed by the measure, the variability produced by these random fluctuations should be considered error variance. While the influence of these psychological states is temporary and random from one time to another, it is quite likely the test-taker will remain in the same state while answering several questions, if not the entire test. Recall an underlying assumption for coefficient alpha is that errors among measurement units are uncorrelated, or, as Becker (2000) notes, "that its violation entails distortion of inconsequential magnitude" (p. 373). If test-takers' psychological state affects their performance from item to item, the errors will be correlated.

Becker (2000) remarked, "there have appeared several articles reintroducing alpha, with special attention drawn to its proper use and concerns for its limitation, assumption violations, and misinterpretations (Cortina, 1993; Miller, 1995; Schmitt, 1996)… yet, the violation of the assumption of uncorrelated errors is mentioned in only one of these three articles, and there it is dismissed as likely being of little import" (p. 373). Following these review articles, a number of investigators began to reexamine the assumptions underlying alpha. The few that looked at the effects of correlated error demonstrated that it can significantly inflate estimates of reliability (e.g., Komaroff, 1997; Raykov, 1998). At first, research in this area was mostly conducted by psychmotericians and statisticians interested in mathematical proofs to their theories. The first application of the research revolved around means to partial out the error variance when correcting observed correlations, a more applied but still surely academic line of research. It wasn't until Becker's work that the issue of transient error and its effect on test score reliability associated with commonly applied measures was systematically examined through an empirical study.

Becker collected self-report ratings of approximately 400 undergraduate university students from three inventories, the Buss-Perry Aggression Question (BPAQ) scale (Buss & Perry, 1992), which contains four scales: Anger, Hostility, Physical Aggression, and Verbal Aggression; the Rosenberg self-esteem (RSE) Scales (Rosenberg, 1965), and the Gender-Free Inventory of Desirable Responding (GFIDR; Becker & Cherny, 1994). Following what he termed a staggered equivalent split-half procedure, error variance associated with transient error was able to be partialed out. Relative to true-score variance the magnitude of transient error was .067 for Physical Aggression, .003 for Verbal Aggression, .021 for Anger, and .145 for Hostility. Transient error was associated with 5.2% of total variance (relative to the estimated true score variance) for the RSE and 11.7% for the GFIDR. Becker concluded that depending on which assumption is more greatly violated, essential τ-equivalence or uncorrelated error, alpha can be a lower or upper boundary of reliability.

While Thorndike (1951) called for investigations toward the influence of transient error (though he did not use the term) half a century before, Becker's (2000) was the first substantive study to demonstrate its influence on the results of applied psychological assessments. Since Becker's work, research on the topic has continued to appear (e.g., Reeve, Heggestad, & George, 2005; Vautier & Jmel, 2003), offering advanced understanding of the ways transient error influences test score interpretation and the best way(s) to assess transient error. For example, while Becker implemented a staggered split-half design, Schmidt, Le, and Ilies (2003) used an approach based upon the calculation of the Coefficient of Equivalence and Stability (CES). While a measure of internal consistency (coefficient of equivalence), such as alpha, can capture variance

associated with random response error and specific factor error, and test-retest measures (measures of stability) can account for random response error and transient error, only the CES can assess error variance from all three sources. Calculating the CES, along with one of the other types of measures, allows the influence of independent error sources to be determined. For example, the difference between the CES and a measure of stability should be due to specific factor error. Alternatively, the difference between the CES and a measure of internal consistency will stem from the influence of transient error. It was in this manner that Schmidt, et al. (2003) conducted their investigation.

Schmidt et al.'s (2003) study replicated and expanded upon Becker's findings. Transient error was found to be present in a variety of commonly used (particularly for personnel selection) measures of individual differences, but the extent to which transient error appeared varied considerably. These measures included the Wonderlic Personnel Test, a test of general mental ability; two separate measures of the Big 5 personality traits (i.e., Contentiousness, Extraversion, Agreeableness, Neuroticism, & Openness to Experience), namely the Personal Characteristic Inventory (PCI; Barrick & Mount, 1995) and the Internations Personality Inventory Pool (IPIP; Goldberg, 1997); Sherer, Maddux, Mercandant, Prentice-Dunn, Jacobs, and Rogers' (1982) Generalized Self-Efficacy Scale (GSE); two measures of self-esteem; and three measures of both positive and negative affectivity. While it was hypothesized that transient error would be smallest in the cognitive domain and largest in areas concerned with affective states (personality traits would fall in the middle depending whether they were more cognitively or affectively loaded), the results were not as straightforward. Transient error associated with the Wonderlic, the cognitive measure, was much less than that associated with positive and

negative affectivity (6.7% versus 17.8% and 14.5%, on average, respectively). The amount was equal to that calculated from the measure of self-efficacy (6.3%) and more than several personality factors (on average: Extraversion was 2.2%, Agreeableness was 3.6%, and Openness to Experience was 0.0%). Schmidt et al (2003) noted: "the primary implication of these findings is that the nearly universal use of the CE (coefficient of equivalence) as the reliability estimate for measures of important and widely used psychological constructs, such as those studied in this research, leads to overestimates of scale reliability" (p. 218).

As can be seen, transient error can only be calculated when a test is completed on two different occasions. Applied practitioners, such as selection analysts, may be left to wonder how these developments should influence their practice, since they rarely deal with data from repeated administrations. Becker called for more research investigating the effects of transient error on various measures in order to determine implications for research and practice. He noted the measures identified as less susceptible to transient errors should be used in place of those that are more liable to be affected by its influence. This is an important point as the alpha point estimate is typically reported without any indication toward the precision of the statistic; the results of two measures that produce similar alpha levels may appear to be equally well suited but the effects of transient error may be much greater for one. Unfortunately, the influence of transient error has only been investigated using a very small number of measures, as the sparse literature review above suggests.

In the interim, one strategy that can be used to account for the overestimation of alpha is to include a confidence interval around the point estimate. Empirical and

conceptual research has demonstrated that the alpha point estimate may lack precision and accuracy in a number of commonly encountered testing situations (Charter & Feldt, 2002) and calls for the presentation of confidence intervals along with point estimates have been made to communicate shortcomings of the statistic. Yet while these calls have been presented in top journals of applied psychological research (e.g., Duhacheck & Iaccobucci, 2004) and educational measurement (e.g., Fan & Thompson, 2001), the impact of these works is not evident. Researchers and practitioners alike may be reluctant to proffer information beyond a reliability coefficient for a variety of reasons. Test developers may fear that the presentation of a lower boundary will negatively affect perceptions of their test, while those who use test results as decision making tools may fear such information will undermine their conclusions. More simply, the rarity of presentation could be due to ignorance regarding their calculation and/or their importance.

Confidence intervals are visible reminders of the fact that a point estimate is not a perfect indicator of scores' true reliability. While this is important for professionals to keep in mind when they present results, it is crucial for the less statistically educated decision makers who use test information to make judgments (as they are more likely to accept the alpha point estimate at its value). Accepting the fact that not only are tests imperfect measures of whatever is being assessed but the statistics used to describe those tests are imperfect indicators as well is critical to the advancement of social science research and the legitimacy of its application. The considerable body of work surrounding corrections to effect sizes is a good example of such advancement. It has been long known (e.g., Johnson, 1944) that one of the effects of measurement error is the

attenuation of obtained effects from reaching the size that would exist if true values, free from error, were obtained. As Baugh (2002) notes, "interpretation of effects without correcting for score unreliability is equivalent to assuming the scores are perfectly reliable even if evidence to the contrary is recognized" (i.e., the reliability coefficient is not 1.00; p. 256). Research has highlighted the common sources of noise that often produce lower effects and methods exist to correct for score unreliability (e.g., Hunter & Schmidt, 1994; Hunter, Schmidt, & Jackson, 1982). Such procedures yield an estimate of the effect size that one might expect to find in a perfect study (Rosenthal, 1994) and thus a true indication of the relation between variables. Correcting effects sizes for unreliability of scores has obvious benefits. As more accurate relations among measures and outcomes are revealed, the utility of those measures will be better understood and, too, their application.

Yet, while comparing the score of one individual to another, such as for selection purposes, reliability information can do little more than provide a general degree of confidence in decisions. Since selection decisions usually stem from a single testing period (though a composite of tests may be used) it is unknown whether individuals' scores on a measure are higher than, lower, or spot on in comparison to their true scores. Thus no corrections can be made to individual scores. So while the reliability of a measure can be used to adjust obtained effects to better understand the true relations among variables, no formulae exist to adjust individuals' scores to better understand the true comparability of their attributes. The standard error of measurement (SEM) statistic does provide test users with some insight toward the accuracy of an individual's test score, but it is not used to adjust obtained scores. While rank-order cannot be changed on

its basis, the standard error of measurement (more specifically the standard error of the difference, a statistic based upon the SEM) can be applied through a technique designed to create ranges of indifference among scores that compensate for the shortcomings of a test's reliability.

Test Score Banding

While test score banding is very commonly applied (for instance, converting percentage correct on a test to a letter grade) a particular form of banding has created a great amount of controversy since its introduction. The technique is known as SED banding, where SED stands for standard error of difference (a statistic related to the reliability of a measure). SED banding (simply referred to as banding heretofore) is based upon the notion that small differences among scores on a test might not be meaningful because the discrepancy could be due to measurement error, as no selection device is perfectly reliable. When scores from an imperfect test are used to make selection decisions among a group of applicants, confidence in those decisions is only as great as the reliability of the measure. Cascio, Outtz, Zedeck, and Goldstein (1991) proposed a technique for creating a band around the highest score on a test so that decision makers can be confident that the scores outside the band are significantly different from the top score. Vice versa, all the scores within the band may not be significantly different from the top score in the band and so are all considered equal. The width of the band is calculated as:

$$Bandwidth = (1.96)\sqrt{2}\ [\ \sigma_x\ (1 - \alpha)\ ] \qquad \text{(Equation 7)}$$

$[\ \sigma_x\ (1 - \alpha)\ ]$ is more commonly known as the standard error of measurement (SEM) and

$\sqrt{2}$ [σ_x (1 - α)] is referred to as the Standard Error of the Difference (SED). "The rationale for specifying the bandwidth as 1.96 X SED is borrowed from the classical hypothesis-testing convention that a null hypothesis should not be rejected if the observed data or more extreme data have at least a .05 probability[4] of occurring if the null hypothesis were true" (when scores are normally distributed; Kehoe & Tenopyr, 1994; p. 297).

SED bands gained popularity because the range of indifference they establish was proposed as an opportunity to create greater opportunity for selecting protected group members. Since many of the most popular types of selection assessments have been demonstrated to produce score differences among Caucasians and protected classes (particularly African Americans, see Table 1), organizations that value diversity often face incompatible choices when it comes to selection measures: they can either choose the test that will yield the greatest return on their investment (e.g., a cognitive ability test; Schmidt & Hunter, 1998) at the expense of diversity goals or they can advance their diversity goals at the expense of their selection device's utility. In an article addressing this issue, Cascio, et al. (1991) introduced several approaches to test scores use that could assist organizations, including several forms of test score banding[5]. While Cascio et al. demonstrated any approach other than strict top-down selection will result in the loss of an assessment's utility, top-down selection can often result in adverse impact against protected classes. If one of the organization's goals is to increase diversity, test score banding could be a viable alternative.

[4] Following a normal distribution of scores, 95% will fall between +/- 1.96 standard deviations.
[5] Several forms of test score banding exist (e.g., fixed vs. sliding); distinctions among them are inconsequential for the purposes of the present research, as any form may be substituted for another.

The process and purpose of the technique is best understood through an example. Imagine there were 50 applicants being selected on the basis of the General Aptitude Test Battery (GATB), a cognitive ability exam. The selection ratio is 10%, i.e. five applicants, test scores ranged from 65 to 95, and the top African American scores were 86, 87, and 89 while five majority applicants scored higher (see Table 2). If selection were conducted top-down, from the highest to the lowest score, African Americans would have no chance of being selected before the selection ratio was met. If a band were created based on an SED equal to 5.00, a band would be created that would roughly range from 85 to 95. Scores above 85 would not be considered significantly different from the top score. Therefore, the top three African American scorers could possibly be selected (depending on the criteria for selection within the band). Proponents of this approach cite such examples as evidence SED banding can help reconcile the differences between an organization choosing a test high in utility and advancing a policy of diversity, through justified scientific means (Zedeck, Outtz, Cascio, & Goldstein, 1991; Cascio, Goldstein, & Outtz, 1995). Critics of SED banding contend the practice is neither scientific nor justified and have questioned the logic and utility of the technique (Schmidt, 1991; Schmidt, & Hunter, 1995).

While theoretical arguments regarding the rationale driving the technique are beyond the scope of the present work, it should be noted that while opponents of banding have made many points that are sound and compelling, the technique has been embraced by practitioners and is not likely to be abandoned any time soon. As such, a practical research course is to determine the most appropriate width to set the band, and thus apply the technique. Guion (2004) provided the following guidance on the topic: " the

reasoning guiding the judgment on the width of the range of indifference should be articulated well enough to be made a matter of written record. This is partly in recognition that business is done in an age of litigation, and one needs to have clear reasoning behind decisions affecting employment processes. (p. 56)."

Using α_{LB} to set the bandwidth

It has been proposed that a confidence interval be presented along with the alpha point estimate in all practical cases. Decision makers will be left to determine how to judge this information, but, hopefully, it will be clear the rank-order of candidates' scores is neither a perfect indication of their knowledge, skills, and/or abilities, nor is it likely to be a perfectly ordinal presentation of the amount candidates will contribute to the organization. While test scores cannot be corrected, ranges of indifference can be created and other factors can be introduced to the selection process. While it is not necessarily advocated here, it is proposed that, based upon the fundamental logic of the SED banding approach, it would be reasonable to substitute the lower boundary of a confidence interval in place of the point estimate when creating a band. This approach would provide test users with the greatest degree of assurance that differences between the band referent (i.e., the top score) and those scores lying outside of the band are truly significantly different. This approach would also create the largest range of indifference and maximize selection opportunities for members of lower scoring groups.

Provided an organization has carefully considered the implications of applying the technique, the reason guiding the band's width (Guion's point) would merely be an extension of Cascio's et al.'s (1991) rationale, which has been accepted in professional

practice. In fact, using the lowery boundary of alpha's confidence interval is logically sounder, as this represents the point of demarcation (i.e., statistically significant difference) based upon the lowest degree of score reliability. The rationale behind the banding technique is to ensure that the top score is significantly different from those scores outside of the band, or those individuals that will not have an opportunity to be selected. A score outside of the band created with Cascio et al.'s (1991) formula may not be statistically significantly different because alpha may be an inflated estimate of score reliability. Using alpha's lower boundary, based upon the calculation of a confidence interval, provides assurance[6] that this does not occur. Thus, Cascio et al.'s calculation would be amedended as such:

$$Bandwidth = (1.96)\ \sqrt{2}\ [\ \sigma_x\ (\ 1 - \alpha_{LB})\ ] \quad \text{(Equation 8)}$$

Where σ = the standard deviation of the test scores and α_{LB} = the lower boundary of alpha's confidence interval

Confidence in Test Results

It is in the interest of personnel practitioners, as well as their clients, to be honest about properties of the tests they create and use. In light of recent research on transient error, it appears coefficient alpha, the favored reliability point estimate most practitioners report, is likely to overestimate the lower boundary of the result's reliability. The presentation of a confidence interval will serve as a reminder that while everything can be done to ensure a test is valid and constructed as well as it possibly can be, there are factors beyond control that will affect the results of any test, though no one can know the full extent of those factors.

[6] Using .05 as the chance level for the reliability estimate and the band.

A unique data set, consisting of test-retest results from an applied setting, provides the opportunity to pull all of these lines of research together. First, the influence of transient error will be demonstrated via a recently proposed method for calculating a reliability estimate from test-retest data. While Becker (2000) and Schmidt, et al. (2003) used derivations of an alternate forms method to calculate transient error, such a design is impractical for researchers interested in applied settings, as this would entail every applicant completing two test administrations. Green (2003) formulated a model that allows transient error to be identified through the analysis of data from the repeated administration of a whole test. A reliability estimate based upon a true-score model with transient error was presented and proposed as a reformulation of coefficient alpha, named test-retest alpha. The coefficient is calculated as:

$$\hat{\alpha}_{X_1X_2} = \frac{k^2\bar{\hat{\sigma}}_{xj_1xj'_2}}{\hat{\sigma}_{x_1}\hat{\sigma}_{x_2}} \quad \text{(Equation 9)}$$

Where k is the number of items on the test, $\hat{\sigma}_{x_1}$ is the standard deviation of the scale from the first administration, $\hat{\sigma}_{x_2}$ is the standard deviation of the scale from the second administration, and $\bar{\hat{\sigma}}_{xj_1xj'_2}$ is the average of pooled different time/different item covariances.

Figure 1 is adopted from Green's (2003) work and illustrates how different sources of variances are captured within and between test administrations. "The test-retest alpha estimates true-score variance based on the different-time/different-item covariances, whereas coefficient alpha estimates true-score variance based on the same-time/different-item covariance" (Green, 2003; p. 89). The presence of transient error can

be empirically demonstrated when test-retest alpha is less than alpha.[7] Green also discusses the difference between test-retest alpha and the test-retest correlation. He notes: "With test-retest correlations, estimates of true-score variance are affected not only by different-time/different-item covariances but also by different-time/same-time covariances. These latter covariances are likely to create an inflated estimate of reliability to the extent that respondents remember how they responded at Time 1 and respond similarly at Time 2" (Green, 2003; p. 89). Green's research was purely mathematical; no empirical data were used to demonstrate his calculations. The present research will empirically demonstrate his work by calculating test-retest alpha and comparing it to both alpha (so transient error can be exposed) and a test-retest Pearson correlation coefficient (so test-retest alpha can be shown as an all-around more precise reliability estimate).

While Green's work will aid applied practitioners who wish for an accurate reliability estimate for the unusual case where they have test-retest data, like Becker's (2000) and Schmidt et al.'s (2003) calculations, transient error cannot be identified from a single test administration. Therefore, the present study echoes calls for the presentation of confidence intervals along with point estimates of reliability, such as by Fan and Thompson (2001). Duhacheck and Iacobucci (2004) recently presented a statistic based on the distribution and standard error of coefficient alpha and demonstrated the superiority of the formula and the confidence intervals created from it in comparison to past calculations (e.g., Feldt & Ankenmann, 1999; Barchard & Hakstian, 1997). Based on the work of van Zyl, Neudecker, and Nel (2000), which presented an asymptotic distribution from the maximum likelihood estimator of the variance of coefficient alpha,

[6] Provided true scores remain constant from the first to second administration.

the authors formulated an estimate of alpha's standard error (ASE). The authors argue that "for the first time, applied and theoretical researchers are able to estimate the standard errors of their measures, thereby revealing precisely the magnitude and severity of the problem of measurement error with less restrictive assumptions on the data" (than past estimates), based upon the ASE (p. 792). For ASE, the distribution of alpha is derived as n → ∞, with $\sqrt{n}(\hat{\alpha}-\alpha)$ following a normal distribution with a mean of zero and a variance of

$$\hat{Q}=\left[\frac{2k^2}{(k-1)^2(j'\hat{V}j)^3}\right]\left[(j'\hat{V}j)(tr\hat{V}^2+tr^2\hat{V})-2(tr\hat{V})(j'\hat{V}^2j)\right] \quad \text{(Equation 10)}$$

Where *n* represents sample size, $\hat{\alpha}$ is the MLE of α, *j* is a k x 1 vector of ones, and *V* is the population covariance matrix among the items (van Zyl et al., 2000). Set with a variance, in the article ASE was derived to equal

$$\text{ASE} = \sqrt{\frac{\hat{Q}}{n}} \quad \text{(Equation 11)}$$

and the appropriate confidence interval (approximately 95%), based on CTT hypothesis testing is

$$\alpha \pm 1.96 \left(\sqrt{\frac{\hat{Q}}{n}}\right) \quad \text{(Equation 12)}$$

Using α_{LB} in personnel selection

While alpha is termed a "point" estimate, the research above (e.g., Becker, 2000; Komaroff, 1997; Schmidt, et al., 2003) demonstrates it is actually a rather blunt instrument for approximating reliability. When other assumptions are met, violation of

essential τ-equivalence deflates alpha as an estimate of reliability, while violation of the uncorrelated errors assumption inflates the statistic. In most applied settings, neither of these assumptions is likely to hold, so the obtained alpha coefficient will not present accurate information about the test results' stability. Presenting alpha with a confidence interval surrounding the point estimate will help better communicate the lack of precision in the measure being used. A likely practical effect of this presentation will be a decrease in the confidence decision makers who use results of the measure to differentiate among test-takers have in their decisions. One of the most likely courses of action personnel practitioners may embrace in such a situation is to create ranges of indifference, based upon the reliability of the test, so decision makers can increase their confidence that true differences do exist between the top scorers and those outside the range of indifference.

The final areas of investigation demonstrate how the confidence interval boundaries surrounding coefficient alpha can be incorporated within a variety of equations to produce more cautious and prudent results. The examples chosen are only a sample of calculations commonly utilized in the practice and research of personnel selection that employ the alpha coefficient. In many cases, using the lower boundary of the reliability estimate can provide practitioners concerned with the inadequacies of their tests, and the statistics used to assess them, a cautious base from which to develop further calculations. As mentioned, one application is to create wider ranges of indifference by substituting α_{LB} in the SED banding equation (Equation 8). The rationale behind this approach is that only when the lower boundary of alpha's confidence interval is used in the creation of the bands can one assert with great confidence (i.e., 95%) that based upon the reliability of the measure the scores outside of the top band are significantly different

(again, at 95%) than those scores that lie outside of the band. The results of this modification are likely to be greater opportunity for minority selections and reduced adverse impact from selection measures.

A second example could be informative for practitioners involved in the test development stage of their selection system. Replacing the traditional point estimate with α_{LB} in the Spearman-Brown prophecy formula (SBPF) will provide test developers a more conservative estimate in determining the effect of increasing the number of items in their test. Namely,

$$\textit{New } \alpha \textit{ level} \quad = \quad \frac{(i)\ \alpha_{LB}}{1+(i-1)\ \alpha_{LB}} \qquad \text{(Equation 13)}$$

Where i = the factor increase in items (e.g., $i = 2$ would be twice the current amount of items)

Take an example where pilot test data shows the internal consistency of a measure to be .67 using α as it is traditionally calculated and the lower boundary of the confidence interval is .57. By doubling the number of items, the test's internal consistency would reach .80 using α, but would only reach .73 using α_{LB} (see Table 3). While there is no uniformly accepted standard of what constitutes high (versus low) levels of reliability, .80 is often used as a threshold. However, the fact that the SBPF using traditional alpha overestimates the SBPF using α_{LB} by 10% should cause some concern. Looking at the example another way, if the test was comprised of 10 items and the developer wished to reach .80 as the level of reliability, using traditional alpha (in the SPBF) would suggest the revised test needs to contain 20 items while substituting α_{LB} would suggest 30 items are necessary to reach the .80 level.

Table 3

Outcome of Spearman-Brown Prophecy Formula Using α and α_{LB}

SBPF Using Traditional α	**SBPF Using α_{LB}**
$\text{New } \alpha = \frac{(i)\ \alpha}{1+(i-1)\ \alpha}$	$\text{New } \alpha = \frac{(i)\ \alpha_{LB}}{1+(i-1)\ \alpha_{LB}}$
$= \frac{(2)\ .67}{1\ +\ (2\ -\ 1)\ .67} = .80$	$= \frac{(2)\ .57}{1\ +\ (2\ -\ 1)\ .57} = .73$

A third and final example could be applied to a variety of contexts in personnel selection practice and, particularly, research. It has become common practice to correct an observed correlation between two variables for attenuation due to measurement error. The general equation is

$$r_{12}^{*} = \frac{r_{12}}{\sqrt{r_{11}r_{22}}} \qquad \text{(Equation 14)}$$

Where r_{12}^{*} = the disattenuated correlation, r_{12} = the observed (attenuated) correlation between variables 1 and 2, and $r_{11}r_{22}$ are the reliability estimates for those variables.

Unlike the previous examples, in this case using α_{LB} will create a more liberal estimate of the correlation between the true scores of the constructs. Take the following example (Table 4) where a predictor measure and a criterion measure have an observed correlation of .70, an alpha point estimate has been found to be .80 for the predictor measure and .80 for the criterion, and a confidence interval around α_{22} has been calculated as +/- .08 (no CI was calculated for the predictor). As can be seen, this

correction can make a notable difference, which of course would be even greater if both the predictor and criterion were corrected for unreliability.

Table 4

Outcome of Correction for Attenuation Formula Using α and α_{LB}

Correction for Attenuation using α_{LB}	Correction for Attenuation using Traditional α
$r_{12}^{*} = \frac{r_{12}}{\sqrt{\alpha_{11}\alpha_{22_{LB}}}}$	$r_{12}^{*} = \frac{r_{12}}{\sqrt{\alpha_{11}\alpha_{22}}}$
$= \frac{.70}{\sqrt{.80(.72)}} = .92$	$= \frac{.70}{\sqrt{.80(.80)}} = .875$

Using α_{LB} in place of the traditional estimate can provide insight into how strong the correlation might be given a "worst case scenario" regarding the reliability of the measure. This could be an effective tool for those revising a measure to be used as a criterion. Comparing newly obtained correlations, using revised editions of the measure, with these estimates can help inform progress.

The present study will use these examples to demonstrate the substantive impact replacing the alpha point estimate with α_{LB} can have on personnel selection practices. The goal of the present research is to draw attention to the shortcomings of the alpha point estimate, present practical applications of recent theoretical advancements in reliability research, and demonstrate just a couple of ways more conservative estimates of reliability can influence the practice of personnel selection.

METHODOLOGY

Sample

The items that make up the composite being reviewed in the present study were part of a larger technical knowledge test on which all candidates who met minimum qualifications for promotion to the rank of sergeant were assessed in both 1999 and 2001. Objections to the results of the 1999 administration were raised in a Federal District Court on the grounds of disparate impact against African Americans. Plaintiffs successfully lobbied for the nullification of the 1999 results and the creation of a new hiring list following the re-administration of the selection procedure, with modifications approved by a Special Master. A Court Order was issued to include various activities leading to revisions of the 1999 selection system and a readministration of the examinations. One hundred and seventy-one police officers (60% Caucasian, 40% African American; 86% Male) completed the same portion of the closed-book examination (with minor "cosmetic" changes; see Appendix A for examples) at both the 1999 and 2001 administrations. These candidates constitute the sample used in the present study.

Measure

In 1999 an external consulting firm was awarded a contract to create and administer several promotional examinations for police positions within a large, municipal Merit System in the Southeastern section of the United States. As a basis for

test development the consultants conducted a comprehensive job analysis for each rank that selections would be made. The consultant's job analysis methodology began with site observations, wherein consultant familiarized themselves with the regular duties performed by incumbents of each rank. Small group interviews followed, in which incumbent subject matter experts (SMEs) generated lists of tasks performed within each rank. After a comprehensive list of tasks was created, panels consisting of larger groups of SMEs were assembled to review the lists and provide additional information that might have been overlooked. A survey was then created, which was composed of all tasks identified by the panels. This survey was disseminated to a large group of incumbents who provided individual ratings about the importance and frequency of each tasks' performance. Based upon results of this survey, tasks deemed critical were identified and concentrated upon in further test development processes.

In a manner similar to the task analysis, areas of knowledge, skills, and abilities (KSAs) that must be possessed in order to effectively perform the tasks established as critical to the sergeant position were identified. First, a sample of incumbents was guided by the consultants to list every area of knowledge, skill, or ability that they used on the job. This list was then transformed into a questionnaire and presented to all incumbents who held the rank. Each sergeant rated the KSA on importance, the frequency of its application, and whether or not a newly appointed sergeant should possess it before assuming the position. Those KSAs that met the threshold for testing were able to be grouped into one of eight categories: Technical Knowledge, Written Expression, Interpersonal Relations, Information Analysis, Judgment and Decision Making, Planning and Organizing, and Resource Management. These categories were tested via three

selection instruments: a Technical Knowledge - Written Test; an In-Basket/Work Sample Test; and an Oral Board Test (see Table 5). As the greatest focus of the current research is placed on the Technical Knowledge component of the selection procedure, only the development of this exam will be further discussed.

The technical knowledge test consisted of two components, an open-book and a closed-book portion. The extent to which each area of knowledge was applied through reference or through recall determined whether it was assessed through open or closed-book testing. Police officers have clearly delineated procedures and protocols for numerous situations. Some cases are obscure and/or not very important and need not be committed to memory. Other areas are critical and/or occur regularly and, as such, need to be committed to memory. The areas of knowledge, for which recall was necessary, identified as most important were administrative practices, state and federal criminal codes, and personnel supervisory practices. For the Sergeant technical knowledge test, items were generated by command-level personnel within the Merit System, with guidance and input from consultants of the external firm. The items were then reviewed by other members of the police command-staff, testing specialists from the consultant's firm, and a linguistic specialist (who reviewed the items for potential biases against protected classes). The final version of the technical knowledge examination consisted of two parts and 103 items; a closed and open-book section, with 59 and 44 items, respectively.

The technical knowledge test and the in-basket/work sample exercise were completed by 391 applicants over a two day period in December of 1999. Three-hundred and one applicants returned to complete the oral board component in early 2000. After

all assessments were scored and a potential hiring list was reviewed, analyses revealed African Americans would be adversely impacted if selections were based on rank-ordered results. Plaintiff parties objected to the results of the examination, the judge ruled in their favor, and no selections were made from the established promotional list. A Court Order was later directed, in May 2001, to redesign and re-administer the examination in accordance with agreed upon modifications.

The consultants contracted to create the 1999 selection process and assessments were retained to complete an updated job analysis to reestablish the tests' content domain and provide opportunities for the plaintiffs, Department of Justice representatives, and other involved parties (e.g., the court appointed a Special Master) to raise objections to material and courses of action during the test (re)design process. The job analysis revealed the duties performed by officers holding the rank of sergeant had not greatly changed since the 1999 analysis. As a result, the examination administered in September 2001 was essentially a replication of the 1999 examination. The three test components used in the 2001 administration were exactly the same as the those used in the 1999 administration in form (i.e., a multiple-choice technical knowledge test, an in-basket/work sample, and an oral board) and function (i.e., presentation and processes followed the same guidelines). The areas of knowledge identified as most critical to performing the duties associated with the position were the same as those for the 1999 administration (Criminal Code, administrative practices, and personnel supervision practices), with the addition of Constitutional Law. Once again, technical knowledge was assessed with a two-part written test, open and closed-book. Only 23 items appeared on the 2001 closed-book technical knowledge test. The presentation of these items on the

2001 administration were nearly identical to their original form on the 1999 administration (see Appendix A for examples of changes).

Three-hundred and sixteen candidates completed the open and closed-book tests of technical knowledge in late September 2001. Of those 316 candidates, 171 had also completed the 1999 technical knowledge test. This group constitutes the sample for the current study. One hundred and three individuals classified themselves as White/NonHispanic, while 68 stated they were African American and the group was predominantly male (N = 148). Of the 23 items that appeared on the 2001 administration of the TK Written Test, 14 had also appeared on the 1999 administration. These 14 items constitute the composite that serves as the main focus of the present study's analyses.

Analyses

Based upon the history of the testing process and the fact that only the items from the 2001 administration were used for hiring decisions, the results from this group items will be the primary focus of analysis. Using the sample and composite described above, coefficient alpha (Equation 4), as it is traditionally calculated, will be computed for the 2001 test to provide a base point of internal consistency. In addition, a test-retest correlation (Pearson) will be produced to indicate the temporal stability of the scores from the first to second administration. Green's (2003) test-retest alpha (Equation 9) will then be calculated and compared to these statistics. The presence of transient error will be exposed if the traditionally calculated alpha (Equation 4) for the 2001 administration is greater than test-retest alpha (Equation 9). Following these analyses, alpha's standard error will be calculated according to Duhachek and Iacobucci's (2004) method, discussed above (Equation 11). Confidence intervals, based on ASE (Equation 12), will be

produced and the location of the various estimates will be presented in relation to the interval.

The upper and lower boundaries of alpha's confidence interval will then be substituted in place of the traditional point estimate for several calculations commonly utilized within the field of personnel selection. First, effects of increasing the test length using both the alpha point estimate and α_{LB} in the Spearman Brown prophecy formula (Equation 13) will be compared. Derivations of the formula will be presented with comparisons being drawn to the results produced by the two statistics. Second, the observed correlation between the technical knowledge composite and each of the other selection measures will be calculated. The observed correlations will be corrected for attenuation due to measurement error (Equation 14) using both traditional alpha and α_{LB}. Additionally, the upper boundary of alpha's confidence interval (α_{UB}) will be substituted in the equation, as the use of this statistic is likely to produce the most conservative estimate. The resulting correlations will be presented and compared.

Finally, comparisons of results applying both the point and lower bound estimates of alpha to Cascio et al.'s (1991) SED banding formula will be presented (Equations 7 & 8, respectively). Differences between bandwidths and the resultant probabilities of minority selections will be demonstrated. Operating under the assumption that within band selections will be made by a (hypothetical) secondary selection device neutral with respect to its effect on racial group membership (i.e., selection rates are random, based on the proportion of each race within the sample), adverse impact analyses will be conducted for each approach (top-down, SED banding, SED α_{LB} banding) and compared.

RESULTS

Coefficient alpha for the 1999 test was calculated to be .333, while coefficient alpha for the 2001 administration was .373. While the level of internal reliability was relatively consistent between the administrations this does not suggest the scores are highly correlated. In fact the correlation coefficient, which shows the temporal consistency between scores on the two administrations, was only moderate (r = .436). Table 6 displays the descriptive statistics for the two tests, while Appendix B presents the covariance matrices for the 1999, 2001, and combined 1999/2001 data.

Table 6

Descriptive Statistics for 1999 and 2001 Administrations

Administration	N	Minimum	Maximum	Mean	Std. Dev.	Alpha
1999	171	5.00	14.00	10.16	1.84	.333
2001	171	7.00	14.00	11.50	1.58	.373

Test-retest Alpha

Green's test-retest alpha was calculated following Equation 9.

$$\hat{\alpha}_{X_1X_2} = \frac{k^2\overline{\hat{\sigma}}_{xj_1xj'_2}}{\hat{\sigma}_{x_1}\hat{\sigma}_{x_2}} = \frac{14^2(.0053)}{1.84(1.58)} = .357$$

Since the test-retest alpha is lower than the traditional alpha calculated for the 2001 test, the presence of transient error could be a factor. Though the difference

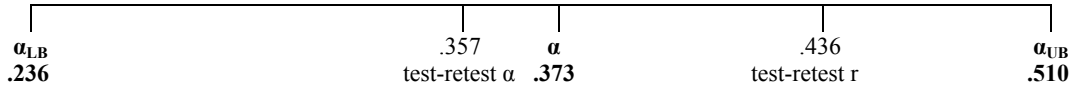
between the statistics is not large, it would not be known to what degree transient error might effect the alpha calculation if the 2001 administration was not the result of the 1999 results being challenged. As such, in order to make the most conservative judgment about the test's reliability and the most cautious application of the statistic in other formulae, the calculation of a confidence interval around the point estimate was produced.

Alpha's Standard Error and Confidence Interval

Alpha's Standard Error (ASE) was calculated to equal .07, following the syntax provided by Duhachek and Iacobucci (2004; located in Appendix C). Using a 95% confidence interval, alpha's lower boundary (α_{LB}) was calculated to equal .236 and alpha's upper boundary (α_{UB}) was calculated to equal .510. Figure 2 clearly shows the confidence interval encapsulates the various reliability estimates.

Figure 2

Confidence Interval and Other Reliability Estimates



Alpha and the Spearman-Brown Prophecy Formula

In cases where the reliability levels are less than desirable (such as the current study) and the researcher would like to know the extent to which increasing the number of items on the measure would improve reliability, the Spearman-Brown prophecy formula can be a useful tool. Table 7 presents a comparison between the estimated effect of doubling the test length (i.e., adding 14 items to the composite investigated in the current study) using both traditional alpha and α_{LB} in the SBPF.

Table 7

SBPF Using Traditional α and α_{LB}

SBPF Using Traditional α	**SBPF Using α_{LB}**
$\text{New } \alpha = \frac{(i)\ \alpha}{1+(i-1)\ \alpha}$	$\text{New } \alpha = \frac{(i)\ \alpha_{LB}}{1+(i-1)\ \alpha_{LB}}$
$= \frac{(2)\ .373}{1\ +\ (2\ -\ 1)\ .373} = .54$	$= \frac{(2)\ .236}{1\ +\ (2\ -\ 1)\ .236} = .38$

As can be seen, the result produced by applying the SBPF using alpha as it is traditionally calculated yields an estimate nearly 1.5 times that of the estimate using α_{LB} (.54 versus .38, respectively). Yet both calculations produce less than desirable levels of internal consistency, so algebraically manipulating the SBPF to determine the number of items necessary to reach such a level, such as .80, could be more informative (see Table 8).

Table 8

SBPF Using Traditional Alpha and α_{LB} (solving for i)

Using Traditional α	**Using Traditional α_{LB}**
$i = \frac{(1-\alpha)\ \alpha_{new}}{(1-\alpha_{new})\ \alpha}$	$i = \frac{(1-\alpha_{LB})\ \alpha_{new}}{(1-\alpha_{new})\ \alpha_{LB}}$
$= \frac{(1-.373)\ .80}{(1-.80)\ .373} = 6.72$	$= \frac{(1-.236)\ .80}{(1-.80)\ .236} = 12.95$

Applying alpha as it is traditionally calculated would suggest that approximately 94 items [6.72 x 14(the original number of items)] would need to be included in order to reach the .80 threshold, while using α_{LB} estimates over 180 items are needed to reach that level of internal consistency. Since it is often unrealistic to create a test with so many items a final calculation is worthwhile to project what the "worst case scenario" of using 94 items might be. Ninety-four items is approximately 6.72 times the amount that originally made up the composite so applying this to the SBPF using α_{LB} figures .67 as the likely ("worst case") level of reliability should the developer decide to use that amount of items.

$$New\ \alpha = (6.72)\ .236 / 1 + (6.72 - 1)\ .236 = .67$$

Alpha and Correction for Attenuation

Another calculation that uses coefficient alpha is the correction for attenuation formula (Equation 14). As mentioned, as part of the promotional testing procedure candidates completed two assessments in addition to the technical knowledge written test, a structured oral interview and an in-basket/role play exercise. The correlation between the 2001 composite and the latter exercise was .203 ($p < .001$) while the correlation with the former was not significant ($r = .015$, $p = .848$). Therefore only the relation between the 2001 composite and the in-basket/role play will be investigated. The alpha level of the exercise (α_{IB}) will remain constant through this example[8] (though corrections for this variable could be appropriate as well) while the upper and lower boundaries of the confidence interval will take the place of the alpha coefficient for the technical knowledge written test (α_{WT}).

[8] While candidates' scores on the in-basket/role play exercise were available their component scores on the exercise were not available. The .67 alpha level comes from the technical manual and represents the whole sample internal consistency of the exercise.

Table 9

Correction for Attenuation using Traditional Alpha, α_{LB}, and α_{UB}

Using Traditional α	$\frac{r_{12}}{\sqrt{\alpha_{IB}\alpha_{WT}}}$	$\frac{.203}{\sqrt{.67\ (.373)}} = .41$
Using α_{LB}	$\frac{r_{12}}{\sqrt{\alpha_{IB}\alpha_{WT_{LB}}}}$	$\frac{.203}{\sqrt{.67\ (.236)}} = .51$
Using α_{UB}	$\frac{r_{12}}{\sqrt{\alpha_{IB}\alpha_{WT_{UB}}}}$	$\frac{.203}{\sqrt{.67\ (.510)}} = .35$

Table 9 shows that correcting the coefficient for attenuation due to measurement error in the composite produces a true correlation of .41 using (traditional) alpha. Substituting α_{LB} instead suggests the linear relation between the two variables could be as high as $\rho = .51$; while inserting α_{UB} in place of coefficient alpha estimates that the variables are only linearly related at $\rho = .35$, after correcting for measurement error.

Alpha and Test Score Banding

Appendix D presents the actual list of candidates, their rank-ordered placement based upon the 2001 composite, and their race. A quick glance at the table shows that a disproportionate number of White candidates are at the top of the distribution. When such results are encountered the SED banding technique may help alleviate the degree of adverse impact associated with the selection process. Table 10 presents the calculation of SED bands using both traditional alpha and α_{LB}.

Table 10

SED Bands using Alpha and α_{LB}

	SED Band	SED α_{LB} Band
Calculation	$(1.96)\ \sqrt{2}\ \left[\sigma_x\ (1-\alpha)\right]$ $2.77\ \left[1.58\ (.627)\right]\ =\ 2.74$	$(1.96)\ \sqrt{2}\ \left[\sigma_x\ (1-\alpha_{LB})\right]$ $2.77\ \left[1.58\ (.764)\right]\ =\ 3.35$
True Band	14.00 – 11.26	14.00 – 10.65
Applicable Band	14.00 – 12.00	14.00 – 11.00

An SED band using coefficient alpha as it is traditionally calculated creates a range of indifference spanning 2.74 points, which translates to a score of 11.26. Since all scores are integers of whole numbers the band includes all scores of 12 and higher, while scores of 11 and lower fall outside of the band. Substituting α_{LB} in place of traditional alpha results in a wider bandwidth. The SED α_{LB} band is equal to 3.35 points, which translates to a score of 10.65. In this case all of those who achieved an 11 or greater on the composite will be included in the band, while those who scored a 10 or lower will not have the opportunity to be selected. For the purposes of the following example, assume a secondary assessment device is employed to make within band selections that result in decisions that are random with respect to race. Table 11 presents the likely number of candidates that would be selected from each racial group following a small (i.e., 7.5%), medium (i.e., 30%) and large (i.e., 55%) selection ratio according to top-down, SED banding, and SED α_{LB} banding selection approaches.

Table 11

Racial Composition of Selected Test-takers by Selection Ratio

Top-down			**SED Band**			**SED α_{LB} Band**		
Select Ratio	Whites	African Americans	Select Ratio	Whites	African Americans	Select Ratio	Whites	African Americans
7.5%	12	1	7.5%	9.5	3.5	7.5%	9	4
30%	42	9	30%	38	13	30%	34.5	16.5
55%	71	24	55%	71	24	55%	64.5	30.5

The level of adverse impact will be contingent upon the approach that is followed. Table 12 presents routinely calculated adverse impact statistics ($4/5^{th}$s Rule calculations and results of Fisher's Exact tests) for each of the three methods. As can be seen, both banding techniques greatly reduce the level of adverse impact associated with the test, though it is still present in many cases. However, substituting α_{LB} in the SED equation creates wider bands that capture more African American candidates, which increases the opportunity of selection and lowers the degree of adverse impact associated with the selection process. While the $4/5^{th}$s rule is still violated in every instance, results of the Fisher Exact tests reveal differences among the three techniques. While the differences in selection rates are not significant at the 7.5% selection ratio using both the traditional and SED α_{LB} bands, SED α_{LB} bands are the only technique that does not result in statistically significant differences at the 30% ratio (while the standard normal deviate is only one hundredth of a point over the adverse impact threshold of 1.96 at the 55% selection ratio).

Table 12

Adverse Impact Calculations by Selection Ratio

Top-down			**SED Band**			**SED α_{LB} Band**		
Select Ratio	$4/5^{th}$s Rule Ratio	Std. Normal Equiv	Select Ratio	$4/5^{th}$s Rule	Std. Normal Equiv	Select Ratio	$4/5^{th}$s Rule	Std. Normal Equiv
7.5%	13%	-2.30	7.5%	45%	-.98	7.5%	67%	-.38
30%	32%	-3.80	30%	52%	-2.35	30%	69%	-1.30
55%	51%	-4.20	55%	51%	-4.20	55%	73%	-1.97

DISCUSSION

Recent research has led to the reexamination of long held assumptions regarding the interpretation of the most commonly presented reliability estimate used in personnel selection research and practice, Cronbach's (1951) Coefficient Alpha. The present study was designed to connect a number of recent advances in this field of research and offer personnel practitioners insight towards the ways these advancements may be applied to their practice. The main area of investigation centered on the effect of transient error, response variations that are due to random changes in test-takers' psychological states across time. A major assumption underlying the alpha coefficient is that it represents the lower boundary of reliability for the results of a measure. Yet, recent theoretical (Komaroff, 1997) and empirical evidence (Becker, 2000) has demonstrated that if errors associated with a measure's items are correlated, the reliability coefficient produced using the alpha calculation can be an inflated estimate of reliability. Transient errors, which are likely to be present in a wide range of testing situations, have the potential to create such a violation. The current study presents the results of a unique data set, where information from an actual selection process yielded the information necessary to identify the influence of transient error. The data also provided the opportunity to demonstrate methods that can protect personnel selection processes against the potential difficulties that result from the influence of transient error.

Test-retest alpha

Several models have been proposed to detect the likely presence of transient error (Becker, 2000; Schmidt, et al., 2003). The current research implemented the test-retest alpha statistic, recently presented by Green (2003), which can reveal the presence of transient error when compared to the traditionally calculated alpha statistic. Because test-retest alpha not only takes within-test/different-item covariances (like traditionally calculated alpha) and between-test/same-item covariances (similar to the test-retest calculations) into account but also the between-test/different-item covariances, this statistic captures all the relevant sources of error assessed through classical test theory. While measures of internal consistency (such as coefficient alpha) can capture variance associated with random response error and specific factor error and measures of stability (such as the test-retest correlation) can account for random response error and transient error, test-retest alpha can account for the variance from all three sources. Transient error, the only source of error not accounted for by alpha, can be identified by subtracting alpha from test-retest alpha. When alpha is larger than test-retest alpha transient error is likely to be inflating the reliability estimate. When alpha is smaller than or equal to test-retest alpha transient error is likely to have a negligible effect on the estimate. Using data from candidates who completed the technical knowledge portion of a promotional examination for the rank of sergeant, in both 1999 and 2001, test-retest alpha was calculated to be .357. Comparing this statistic to the alpha calculation for the composite from 2001 (.373) and the test-retest correlation between the 1999 and 2001 administrations (.436) reveals test-retest alpha is the lowest estimate among the three. This suggests transient error may have inflated alpha as an estimate of reliability for the

2001 administration.

Though the example provided in the current study demonstrates transient error can be a factor that effects the calculation of coefficient alpha, the difference is small (.016). While it might be easy to dismiss such a small discrepancy as an insignificant factor that would not impact the way the test is viewed, this is only a single sample and should not suggest that transient error in innocuous. There is no particular value where inflation will impact interpretation of test results. In some cases a few hundredths of difference might be influential, while in others a couple tenths could have no practical effect. The lack of concrete guidelines to gage the influence of transient error may discourage practitioners who wish to control for its effects, but the problem is quite insignificant considering the fact this source of error can almost never be identified in most testing situations. Practitioners who create and analyze the results of tests from a single administration would not have the data to generate test-retest alpha, or any other statistic that can be used for similar purposes (e.g., Becker, 2000, Schmidt, et al., 2001), thus the extent to which alpha is being over- or underestimated cannot usually be known. To compensate for this difficulty the present study calls for the confidence interval to be presented and used more regularly in personnel selection research and practice.

Confidence interval alpha

In personnel selection contexts the validation of tests are often demonstrated through a content-based approach, where documentation of incumbent subject mater expert ratings of the test material serves as evidence that the assessment is appropriate. This is in contrast to the statistical information that is produced following construct and/or criterion-related validation approaches. The absence of additional statistical information

places a greater weight on the reliability estimate as a diagnostic instrument. This lone statistic could be the only factor used to interpret candidates' scores and make selection decisions. With so much emphasis placed on a single estimate it is imperative testing professionals communicate the degree of precision associated with the calculation. While the alpha point estimate is often the best single estimate to consider when interpreting the reliability of results from a test, additional diagnostic information is readily available to be presented along with the statistic.

The present study echoes the call of recent researchers (e.g., Duhachek & Iacobucci, 2004) to supplement the alpha point estimate with additional information, such as the standard error of the calculation and a confidence interval. Adding this information generates a visible reminder that coefficient alpha is not a perfect indicator of test scores' true reliability. In the current study the upper and lower boundaries of alpha were computed using the statistic and formula developed by Duhachek and Iaccobucci (2004). Alpha's Standard Error (ASE), which was found to equal .07, is based upon the distribution of standard error surrounding coefficient alpha and serves as the basis for creating the confidence interval. The resultant confidence interval, which ranges from .236 (α_{LB}) to .510 (α_{UB}), suggests that if 100 samples were taken of this composite, the alpha coefficient would be calculated to fall within those upper and lower boundaries 95 times out of 100. It is suggested that this information cannot only be influential in decision-makers' interpretation of test results but can also help inform personnel practitioners in creating and utilizing the tests they develop.

Utilizing the Confidence Interval

Beyond its use as a measurement for the level of internal consistency, coefficient alpha also serves as the anchor statistic in a variety of formulae used to project, interpret, and apply test results. If the estimate is inaccurate, so too are the results of whatever formula used the estimate in its calculation, which in turn could lead to inaccurate interpretations and applications. In high stakes settings, such as personnel selection, it is prudent to err on the side of caution. If the possibility exists that alpha overestimates the reliability of a set of results, because of the effect of transient error, and calculations exist to formulate a more conservative statistic, caution can and should be exercised. The present study demonstrates the benefits of substituting the upper and lower boundaries of alpha's confidence interval in place of the point estimate in a variety of calculations common to personnel selection.

First, the Spearman-Brown prophecy formula, which informs researchers of the extent to which increasing the number of items on their measure would improve reliability, was used as an example to demonstrate the effects of substituting α_{LB} in place of alpha as it is traditionally calculated. The results were quite informative. As Table 7 presents, when projecting the reliability level that results from doubling the current number of items on the composite, the result produced by applying the SBPF using alpha as it is traditionally calculated yields an estimate nearly 1.5 times that of the estimate using α_{LB}. Applying alpha as it is traditionally calculated would suggest that doubling the test length from 14 to 28 items would improve reliability from .37 to .54 (a 46% increase), while substituting α_{LB} in the equation suggests the improvement would be one hundredth of a point, from .37 to .38 (less than a 3% increase). Basically, the result of

the α_{LB} substitution suggests the test length would need to be doubled in order to assure the new reliability level will not be less than the original point estimate.

Manipulating the SBPF to calculate the number of items necessary to reach an "acceptable" level of reliability, in the present case .80, also provides a marked difference between alpha and α_{LB}. Using alpha as it is traditionally calculated suggests approximately 94 items would need to be included in order to reach the .80 threshold, while using α_{LB} estimates over 180 items would be needed to reach that level of internal consistency. Since it is most likely unrealistic to create a test with so many items, it was demonstrated that other derivations of the SBPF can utilize the α_{LB} substitution to provide additional information, such as what the "worst case scenario" of using a certain amount of items on a test might be. Using the current example, 94 items was determined as the number of items necessary to reach an alpha level of .80, when alpha as it is traditionally calculated was applied to the formula. Since 94 items is 6.72 times the original amount, applying this number to the SBPF using α_{LB} estimates .67 as the projected level of reliability. A researcher who conducted such calculations could then be reasonably sure the alpha level that will result from including 94 items on the test will not be below .67., though it is more likely to be near .80.

The results of the current study demonstrate that the discrepancy between results of the SBPF using alpha as it is traditionally calculated and α_{LB} can be significant, but whether the SBPF over- or underestimates the new reliability level will depend on which assumptions underlying the alpha calculation are violated (essential tau-equivalence and/or uncorrelated error). Of course finding a higher than expected level of reliability is not undesirable, but a lower than expected result could have severe implications for test

development planning. For example, a researcher could develop a test, conduct a pilot study, calculate the alpha level of results, perform the SBPF calculation to determine the effect of doubling the amount items that appeared on the pilot version, create the new items, re-administer the test, and find a completely different reliability level than predicted by the traditional formula.

To demonstrate the practical effect of overestimating reliability predictions consider the following. The test-retest data used in the present study was the product of a successful court challenge to the results of the 1999 administration. The consequence was a hiring freeze where no officers were promoted to the rank of sergeant after the 1999 test. The two year period during which no promotions were made undoubtedly placed a greater strain on the existing group of sergeants. The effect of the additional strain could have led to a decrease in the sergeants' performance, which in turn could have led to decreases in arrests, convictions, and/or increases in crime. Although it was not so in the present case, an unacceptable level of reliability can serve as grounds for challenging the results of a selection process. If one of the challenges brought against the results of the 1999 test was a low level of reliability and the researcher increased the number of items on the 2001 administration to reach a certain level of internal consistency based upon the SBPF using alpha as it is traditionally calculated, an unexpectedly low alpha level could have once again been obtained providing additional grounds for challenge.

Although operating under a court ordered consent decree is an atypical situation for most organizations and serious consequences, such as increased crime rates, might seem far removed from the calculation of a reliability estimate, the degrees of separation

between imprecise calculations and real world negative effects are actually quite small. There is path that begins with test interpretations and ends with customer impressions. A reliability estimate informs users of a selection process' value, while the selection process informs organizations of the extent to which the candidates they choose will perform effectively, and the candidates who are hired to perform for the organization determine the worth of products or services provided to the public. Most business models will plot a straight course that leads to the goal of providing customers with products or services that are well received, but if the first step is askew the path will be off-mark. Since no models, tests, or statistics are perfectly accurate, a conservative approach should be adopted to guard against over-projections (again, exceeding projections is not as undesirable). Test development procedures based upon cautious calculations, such as substituting α_{LB} in the SBPF, allow personnel practitioners to protect the organizations they serve against less than desirable test results, which will help assure the quality of all subsequent decisions and successive outcomes.

Correction for Attenuation

The second example of utilizing the boundaries of alpha's confidence interval focused on the role of the reliability coefficient in correcting correlation coefficients for attenuation due to measurement error. Only the results from the in-basket/work sample exercise resulted in a significant correlation with the composite, though it was a low correlation (.206). Table 9 presents the results of correcting the correlation coefficient using alpha, α_{LB}, and α_{UB} and, once again, the results are notable. Using alpha as it is traditionally calculated resulted in a revised estimate of the correlation equaling .41, while substituting α_{LB} produced a coefficient that was approximately 25% higher (.51).

While substituting α_{LB} could be warranted if the researchers knew that transient error was inflating the estimate, as mentioned, it is rather unusual to have the test-retest data necessary to make this determination. If α_{LB} is used without this information the "corrected" correlation coefficient could be a severely inflated estimate. A more cautious approach would be to instead apply α_{UB}. In the current example, the corrected correlation coefficient using α_{LB} was calculated to equal .35. Obtaining all three estimates, the conservative estimate using α_{UB} (.35), the primary figure using the alpha point estimate (.41), and the liberal estimate using α_{LB} (.51) would, of course, be the most informative.

The additional information provided by substituting the confidence interval boundaries in place of the point estimate when correcting correlation coefficients for attenuation due to measurement error could be resourceful for test developers engaged in establishing the construct validity of a new measure, for researchers comparing results among studies, and for authors presenting results of newly investigated relations among variables. The purpose of introducing this example, as well as the SBPF example, is to demonstrate that the lack of precision in reliability estimates should not remain a merely theoretical interest. Reliability estimates are included in a great number of commonly applied formulae, which effect the ways tests are created and utilized and their results interpreted. The final area of investigation may best demonstrate the degree to which reliability estimates can exert salient effects on personnel selection decisions.

SED Banding

The topic of SED banding has a very divisive history. Though there are a number of strong supporters who have tirelessly endorsed the practice in the professional literature and an almost consensus opinion that the creation of bands in and of themselves

(that is without discussion of band width or the ways within band selections are made) is a legally defensible practice, a strong group of critics also exist who challenge the use of the technique. Many of the criticisms are directed at the theoretical and logical grounds that serve as the technique's basis. At the core of those grounds is the argument that it is reasonable to create ranges of indifference among scores because no test is a perfect measurement device. The proof of this assertion is the fact that reliability coefficients are almost never equal to 1.00. While this fact provides a very strong argument for the general strategy of banding, the specific mechanics of the approach, i.e. determining a range of indifference, must also be defended.

As Equation 8 demonstrates, along with the standard deviation of the test scores, the bandwidth will differ from test to test as larger and smaller reliability coefficients, which are routinely coefficient alpha, are inserted into the formula. If the purpose of the creating the band is to assure those scores that lie outside the band are significantly different (based upon a predetermined level of confidence) from those within the band, precision of the alpha coefficient is paramount. If the alpha coefficient is an underestimate, the bandwidth will be larger than it needs to be. Thus, some of the lowest scores incorporated within the band should actually be left outside of it. If the alpha coefficient is an overestimate, the bandwidth will not be large enough. In this case, some of the highest scores that lie just outside of the band should actually be incorporated within it.

Incorrectly stating a score outside of the band is significantly different from the top score within the band is similar to conducting a Type I error in research, while including a score within the band that should actually remain on the other side is similar

to creating a Type II error. The former will occur when an alpha coefficient that overestimates reliability is included in the SED band equation. In such cases individuals who lie just outside of the band are treated as being significantly different from the referent score (similar to rejecting H_O) though they are not (statistically significantly different based upon the prescribed confidence level). Conversely, applying an alpha coefficient that underestimates reliability is comparable to a Type II error. Here some of the individuals within the band are significantly different from the referent scorer, though they are treated as being equivalent (similar to retaining H_O).

In personnel selection settings, just as any other, a choice must be made whether to favor committing a Type I or a Type II error. The decision to use a potentially wider than needed band (following a Type II error) over a potentially narrower band (following a Type I error) should be heavily influenced by the repercussions from each choice. For instance, while the owner of a new retail store may comfortably commit a Type II error when selecting security guards, because the greatest repercussion will likely be small amounts of shoplifting or loitering teenagers, the owner of a new sightseeing service should not feel as comfortable committing this type of error when selecting pilots to carry clients, because the repercussions in this case include loss of life and extremely expensive equipment. Since most scenarios lie somewhere between these extreme examples the choices are usually more difficult and how well employees perform their job is only one criterion to consider when making a hiring or promotion decision. There are a host of other considerations that should be weighed such as employee development and career progression, public image and relations, as well as diversity and affirmative action goals.

Generally, creating wider bandwidths provides greater opportunity for minority selections, though this depends upon the degree to which the secondary selection device (used for within band selections) effects the selection rates of minority groups. The present study demonstrates that when a secondary selection device has a random effect on racial selections, the wider the band the greater the opportunity for minority selections. While using an SED band as it is traditionally calculated (using the alpha point estimate) would result in a marked increase in minority selections over strict top-down selections (particularly with smaller selection ratios), expanding the band by substituting α_{LB} in place of the alpha point estimate creates even greater opportunity. Table 12 demonstrates the impact these techniques have on adverse impact calculations. While the 4/5ths Rule remains violated following every technique at every selection ratio, the proportion of minority to majority selections greatly improves with the creation of traditional SED bands, and to an even greater extent using SED α_{LB} bands. Looking at the 30% selection ratio, the 4/5th ratio using top-down selection is 32%, but increases to 52% after applying the SED band (a 162% improvement) and 69% after applying the SED α_{LB} band (a 216% improvement). The results are even more dramatic for the smaller selection ratio (7.5%), where a 13% minority to majority selection ratio exists following top-down selections. In this case, using an SED band improves that ratio over 300% (4/5ths calculation = 45%), while an SED α_{LB} band improves upon the top-down ratio by over 500% (4/5ths calculation = 67%).

As 4/5ths Rule calculations are heavily dependent upon sample size the best professional practice calls for a statistical significance test to supplement the figure. In the present study the Fisher Exact test was employed to test the null hypothesis that no

significant differences exist between the selection rates of minority and majority applicants. Table 12 presents the results of these tests as well. When the standard normal equivalent exceeds 1.96 the selection rates between the two groups are considered to be statistically significantly different. As can be seen, following top down selection at the 7.5% selection ratio produces a result that crosses this threshold, demonstrating there is a statistically significant difference between the selection rates of African Americans and Whites. When both the SED band and the SED α_{LB} band are applied the standard normal equivalent drops below one, revealing the difference between selection rates for the two groups could be due to chance. At the 7.5% selection ratio both the SED and SED α_{LB} bands would be successful means to combat adverse impact (as detected by the Fisher's Exact test). However, at the 30% selection ratio only the SED α_{LB} band would be effective in this manner. At the 55% ratio note the SED band produces a result equal to the top-down procedure, a statistically significant standard normal equivalent of -4.20, while the SED α_{LB} band just barely crosses the 1.96 threshold. The results of the present study clearly demonstrate that substituting the lower bound alpha in place of the alpha point estimate expands SED bands to a point where a substantive improvement can be observed when assessing adverse impact.

While the α_{LB} substitution in the SED banding formula can be seen as one of many ways the confidence interval created around alpha can be utilized by personnel practitioners, it is quite different from most other examples and thus deserves further discussion. Unlike the correction for attenuation and Spearman-Brown prophecy formula examples, where the α_{LB} substitution can be performed solely for informational purposes, applying the SED banding modification proposed here will very often have immediate

real world effects (e.g., some individuals will be given the opportunity for promotion while others will not). Given the amount of controversy that surrounds the SED banding technique in general, the question of how well this modification will be accepted by the community of personnel practitioners should be addressed.

Judging the Use of α_{LB} as a Professional Practice

There are three main criteria for acceptance upon which the approach will likely be judged. The first and foremost issue to be weighed is the logic underlying the modification. Is it reasonable to substitute the lower boundary from the confidence interval surrounding alpha in place of the point estimate? Based upon recent research that has demonstrated transient errors can cause a violation of the uncorrelated errors assumption and lead to an inflated reliability estimate (via coefficient alpha) in almost any testing situation, the modification to the SED banding technique is not only intuitively reasonable but fully in line with the logic and purpose of applying the technique. SED bands are formed to ensure individuals with the requisite knowledge, skills, and abilities for successfully performing the duties of the position being tested are not passed over due to the presence of measurement error in the instrument used to make the assessment. Due to the manner in which the bandwidth is calculated (i.e., using variability and reliability as the major determinants), if alpha were overestimated the bands would be narrower than intended. This outcome may in turn create a situation wherein individuals with true levels of competency equal to the individual with the highest score on the test may not be given the opportunity to be selected not because of his or her true level of knowledge, skills, or abilities but due to error in measuring those traits. The use of α_{LB} helps makes certain this does not occur by creating a band with

confidence that the reliability estimate used in the equation is not inflated.

The second criterion against which practitioners will likely weigh the modification is the efficacy of the technique. While it is reasonable to use the revision on the bases of logic and reason, it is likely practitioners would only be interested in adopting it if it were demonstrated to assist organizations in improving their diversity and/or reducing adverse impact. The present research supports the notion that SED α_{LB} bands can indeed lead to these outcomes. Both the $4/5^{th}$ calculations as well as the results from the Fisher Exact tests demonstrate that SED α_{LB} bands produce greater opportunity for minority selection, which reduces the likelihood of adverse impact and increases the potential for a diverse organization. Of course, like all banding techniques, the utility of the selection procedure will decrease, though the modification will likely lead to a greater decrease than that resulting from traditionally calculated SED bands. As discussed, a decision must be made whether to favor the integrity of the selection procedures or the advancement of diversity goals. If an organization is determined to choose the latter, SED α_{LB} banding represents a logically sound alternative that can improve the probability of protected class selections. However, future research should investigate the tradeoff in utility between traditionally calculated SED bands and the proposed modification so that practitioners can be fully informed when assisting an organization with the decision.

The third and final criterion is unfortunately the most difficult to assess, that is the likelihood the modification will be accepted by the courts. An in-depth discussion of this matter is beyond the scope of the present work, but as the acceptability of the modification is inherently linked to the acceptability of the SED banding technique in general, previously published literature covering this topic may be instructive (e.g.,

Barrett & Lueke, 2004). Since the approach has a greater likelihood to assist minority groups, discrimination lawsuits initiated by protected classes in opposition to the technique would be less likely to arise than those following top-down selection, or even those following traditionally calculated bands. However, the chances that charges of reverse discrimination being filed by white applicants with the top rank-ordered scores could increase beyond those encountered using both top-down selection and traditionally calculated SED bands; the wider the band the greater the potential for this type of challenge. Therefore Guion's (2004) advice about determining the width of the range of indifference using reason that can be "articulated well enough to be made a matter of written record" (p. 56) becomes a critical issue.

Fortunately the argument for using SED α_{LB} banding is uncomplicated. There are no perfect tests; all contain measurement error that affects interpretation of results, especially differences among scores. In a context as important as personnel selection, the accuracy of decisions to favor some individuals over others is crucial. If decision makers cannot be confident about the distinctions they would like to draw, creating a range of indifference, wherein all scores are considered equal, is a reasonable means of compensation. The SED banding technique calculates this range using the variability and reliability of the obtained results (which are at the root of the uncertainty in the results). In order to exercise the greatest amount of caution while drawing distinctions among applicants, the bandwidth is determined by employing the most conservative estimate of reliability that is available, α_{LB}. Though the technique could still be attacked for the same reasons traditional SED banding has been criticized (e.g., it does not represent the "best practice" for selection procedure utility), the modification should not introduce any new

data are available. The only practicable alternative in most cases is to compensate for the imprecision of the coefficient. The most direct way to accomplish this is to offer a confidence interval that surrounds the alpha point estimate. Presenting this supplemental information is the best means to affect decision makers' understanding of the statistic, thereby influencing interpretation of the test results they have obtained and the personnel decisions they must conclude.

The upper and lower boundaries of alpha's confidence interval not only provide valuable information to be used in the interpretation of the statistic but are sound estimates of reliability in their own right that researchers and practitioners can apply to other formulae they commonly employ. The examples provided in the present study demonstrate the advantages of substituting the confidence interval boundaries in place of the point estimate. These advantages stem from practicing caution in application and being conservative in interpretation. Application of the social sciences, such as the design of personnel selection procedures, can never be conducted with the same degree of accuracy as the more concrete, natural sciences. Yet, so long as the degree and sources of inaccuracy are never concealed, advancements will continue and shortcomings can be minimized.

obstacles to judicial acceptance.

Conclusion

The overarching theme of the present research is conservatism. One of the reasons coefficient alpha has been so widely applied and accepted is due to the long held notion that the calculation presents the most conservative estimate of reliability. However, in light of recent research that has demonstrated factors exist (i.e. transient error) that can cause coefficient alpha to present an inflated estimate of reliability, professionals must reassess the meaning they assign to the venerable statistic and the influence it has on their practice. Three logical alternatives exist: 1) coefficient alpha can be abandoned for other reliability estimates that better account for these factors, 2) corrections can be made to the calculation to prevent the overestimation, and 3) compensatory techniques can be employed to offset potential shortcomings when applying and interpreting the statistic. Though the first alternative is certainly a viable option, coefficient alpha is so widely used and recognized that its abandonment could have negative repercussions on a wide array of research and practice. For example, while reliability estimates following Generalizability Theory would be appropriate (and much more informative) in many personnel selection contexts, for better or worse a significant amount of personnel practitioners are not as familiar with the methods of conducting a g-study, and even fewer decision makers within the organizations they serve could well interpret the results meaning.

Unfortunately, the means to correct the calculation are usually unavailable, leaving the second alternative an unviable option as well. Test-retest alpha is an example of a "corrected" coefficient alpha but the statistic can only be computed when test-retest

REFERENCES

Americans with Disabilities Act of 1990. 1990. P. L. 101-336.

Barchard, K.A. & Hakstian, A.R. (1997). The robustness of confidence intervals for coefficient alpha under violation of the assumption of essential parallelism. *Multivariate Behavioural Research, 32*(2), 169-191.

Barrett, G.V. & Lueke, S.B. (2004). Legal and practical implications of banding for personnel selection. In H. Auginis (Ed.). *Test-score banding in human resource selection: Technical, legal, and societal issues.* Praeger Publishers/Greenwood Publishing Group, Westport, CT.

Barrick, M. R., & Mount, M. K. (1995). The Personal characteristics inventory manual. Unpublished manuscript, University of Iowa, Iowa City.

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement, 62,* 254-263.

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods, 5,* 370 – 379.

Becker, G., & Cherny, S. S. (1994). Gender-controlled measures of socially desirable responding. *Journal of Clinical Psychology, 50,* 746-752.

Buss, A. H., & Perry, M. (1992). The Aggression Questionnaire. *Journal of Personality and Social Psychology, 63,* 452-459.

Campion, M. A., Outtz, J. L., Zedeck, S., Schmidt, F. L., Kehoe, J. F., Murphy, K. R., & Guion, R. M. (2001). The controversy over score banding in personnel selection: answers to 10 key questions. *Personnel Psychology, 54,* 149-185.

Cascio, W. F., Goldstein, I. L., & Outtz, J. (1995). Twenty issues and answers about sliding bands. *Human Performance, 8,* 227-242.

Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4,* 233-264.

Charter, R. A., & Feldt, L. S. (2002). The importance of reliability as it relates to true score confidence intervals. *Measurement and Evaluation in Counseling and Development, 35,* 104-112.

Civil Rights Act of 1991, 42 U.S.C. §2000e-2(1)

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Crano, W. D., & Brewer, M. B. (1973). *Principles of research in social psychology*. in Crocker, L., & Algina, A. (Eds.) (1986). *Introduction to classical and modern test theory.* Harcourt Brace Jovanovich College Publishers, New York.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Harcourt Brace Jovanovich College Publishers, New York.

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

Duhachek, A., Coughlan, A. T., & Iacobucci, D. (2005). Results on the standard error of the coefficient alpha index of reliability. *Marketing Science, 24,* 294-301.

Duhachek, A. & Iacobucci, D. (2004). Alpha's standard error (ASE): an accurate and precise confidence interval estimate. *Journal of Applied Psychology, 89,* 792-808.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement, 61,* 517-531.

Feldt, L. S., & Ankenmann, R. D. (1999). Determining sample size for a test of the equality of alpha coefficients when the number of part-tests is small. *Psychological Methods, 4,* 366- 377.

Fiske, D. W. (1966). Some hypotheses concerning test adequacy. *Educational and Psychological Measurement, 26,* 69-88.

Goldberg, L. R. (1997). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F.De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7-28). Tilburg, the Netherlands: Tilburg University Press.

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods, 8,* 88-101.

Guion, R.M. (2004). Banding: Background and general management purpose. In Aguinis, H. (ed.), *Test-score banding in human resource selection: Technical, legal, and societal issues.* Praeger Publishers: Westport, CT.

Hakstian, A.R., & Whalen, T.E. (1976). A K-sample significance test for independent alpha coefficients. *Psychometrika*, 41, 219-231.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Hoyt, C.J. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, *6,* 153-160.

Hunter, J. E., & Schmidt, F. L. (1994). The estimation of sampling error variance in the meta-analysis of correlations: Use of r in the homogenous case. *Journal of Applied Psychology, 78,* 171-177.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Newbury Park, CA: Sage.

Johnson, H. G. (1944). An empirical study of the influences of errors of measurement upon correlation. *American Journal of Psychology, 57,* 521-536.

Kehoe, J. F., & Tenopyr, M. L. (1994). Adjustment in assessment scores and their usage: A taxonomy and evaluation of methods. *Psychological Assessment, 6,* 291-303.

Knapp, T. R. (1991). Coefficient alpha: conceptualizations and anomalies. *Research in Nursing and Health, 14,* 457-460.

Komaroff, E. (1997). Effect of simultaneous violations of essential τ-equivalence and uncorrelated error on coefficient α. *Applied Psychological Measurement, 21,* 337-348.

Komaroff, E. (1997). *SEMNET discussion on alpha and correlated error.* Retrieved February 10, 2006 from http://bama.ua.edu/cgi bin/wa?A2=ind9710&L=semnet&T=0&F=&S=&P=770

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika, 39,* 491-499.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Miller, M. B. (1995). Coefficient alpha: a basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2,* 255-273.

Murphy, K. R., & Davidshofer, C. (2001). *Psychological Testing: Principles and Applications* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32,* 1-13.

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogenous items. *Applied Psychological Measurement, 22,* 35-385.

Reeve, C. L., Heggestad, E. D., & George, E. (2005). Estimation of transient error in cognitive ability scales. *International Journal of Selection and Assessment, 13,* 316 – 320.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation

Schmidt, F. L. (1991). Why all banding procedures in personnel selection are logically flawed. *Human Performance, 4,* 265-277.

Schmidt, F. L., & Hunter, J. E. (1995). The fatal internal contradiction in banding: its statistical rationale is logically inconsistent with its operational procedures. *Human Performance, 8,* 203-214.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: an empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8,* 206-224.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. Psychological Methods, 8, 350-353.

Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The self-efficacy scale. *Psychological Reports, 76,* 707-710.

Society for Industrial and Organizational Psychology. (2003). Principles for the validation and use of personnel selection procedures. Bowling Green, OH: SIOP.

Thompson, B. (2003). *Score Reliability: Contemporary Thinking on Reliability Issues.* Thousand Oaks, CA: Sage Publications.

Thorndike, R.L. (1951). Reliability. In Lindquist, E.F. (ed.), *Educational Measurement.* ACE, Washington, DC.

Uniform Guidelines on Employee Selection Procedures, 43 Fed. Reg. 38290-38315 (1978).

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65,* 271-280.

Vautier, S., & Jmel, S. (2003). Transient Error of Specificity? An alternative to the staggered equivalent split-half procedure. *Psychological Methods, 8,* 225-238.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Zedeck, S., Outtz, J., Cascio, W. F., Goldstein, I. L. (1991). Why do "testing experts" have such limited vision? *Human Performance, 4,* 297-308.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53,* 33-49.

APPENDICES

Appendix A

Example Item Change from 1999 Administration to 2001 Administration

1999 - When executing an arrest warrant, there are several procedures which must be carefully followed to ensure that the warrant is executed properly. For example, if the arrestee demands to see the warrant before the arrest is made, the arresting officer must:

a) show the warrant to the arrestee and allow the arrestee to examine it, before proceeding with the arrest.

b) show the warrant to the arrestee as soon as practicable, even if that time is after the arrest.*

c) explain the cause of the arrest either by stating the substance of the warrant or by reading it to the arrestee.

d) issue a copy of the warrant to the arrestee as soon as practicable, even if that time is after the arrest.

2001 - If an arrestee demands that an officer executing an arrest warrant show the warrant before the arrest is made, the arresting officer must:

a) explain the cause of the arrest by reading the warrant to the arrestee before proceeding with the arrest.

b) provide the arrestee with a copy of the warrant as soon as practicable even if that time is after the arrest.*

c) show the warrant to the arrestee as soon as practicable even if that time is after the arrest.

d) allow the arrestee to examine the warrant before proceeding with the arrest.

Appendix B

COVARIANCE MATRICES

1999 Covariance Matrix

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14
Item 1	.090	.007	-.002	.010	.015	.023	-.001	.004	-.016	.011	-.014	.008	.004	.005
Item 2	.007	.250	-.009	-.019	.012	-.008	-.003	.012	-.003	-.011	.015	.001	-.004	.018
Item 3	-.002	-.009	.189	.010	.017	.032	-.001	.014	.025	.016	.009	.004	-.009	.007
Item 4	.010	-.019	.010	.246	.028	.015	.003	.008	.041	-.018	.003	.016	.002	.008
Item 5	.015	.012	.017	.028	.121	.037	-.001	.024	.008	-.021	.013	.004	.027	.009
Item 6	.023	-.008	.032	.015	.037	.232	-.002	.010	.006	-.001	.033	-.019	.007	.018
Item 7	-.001	-.003	-.001	.003	-.001	-.002	.006	.002	.004	-.002	.005	-.001	-.003	.000
Item 8	.004	.012	.090	.007	-.002	.010	.015	.023	-.001	.004	-.016	.011	-.014	.008
Item 9	-.016	-.003	.007	.250	-.009	-.019	.012	-.008	-.003	.012	-.003	-.011	.015	.001
Item 10	.011	-.011	-.002	-.009	.189	.010	.017	.032	-.001	.014	.025	.016	.009	.004
Item 11	-.014	.015	.010	-.019	.010	.246	.028	.015	.003	.008	.041	-.018	.003	.016
Item 12	.008	.001	.015	.012	.017	.028	.121	.037	-.001	.024	.008	-.021	.013	.004
Item 13	.004	-.004	.023	-.008	.032	.015	.037	.232	-.002	.010	.006	-.001	.033	-.019
Item 14	.005	.018	-.001	-.003	-.001	.003	-.001	-.002	.006	.002	.004	-.002	.005	-.001

2001 Covariance Matrix

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14
Item 1	.241	-.014	.001	.001	.004	.018	-.002	-.028	-.014	-.007	.028	-.017	-.033	-.009
Item 2	-.014	.251	.007	.025	.010	.005	.003	.025	-.002	.003	.007	.004	.009	-.006
Item 3	.001	.007	.113	.025	.008	.013	-.001	-.007	.002	.004	.004	.010	.00	-.003
Item 4	.001	.025	.025	.113	.008	.001	-.001	.005	-.010	-.002	.010	-.007	.012	.003
Item 5	.004	.010	.008	.008	.108	.026	-.001	.006	.002	.004	.005	.011	.026	.009
Item 6	.018	.005	.013	.001	.026	.195	-.002	.009	.015	-.005	.007	.020	.004	.006
Item 7	-.002	.003	-.001	-.001	-.001	-.002	.006	-.001	-.001	.000	-.001	.005	-.002	.000
Item 8	-.028	.025	-.007	.005	.006	.009	-.001	.153	.018	.003	-.002	.008	.044	.007
Item 9	-.014	-.002	.002	-.010	.002	.015	-.001	.018	.108	.004	-.013	.023	.014	.003
Item 10	-.007	.003	.004	-.002	.004	-.005	.000	.003	.004	.017	-.002	.003	.005	.000
Item 11	.028	.007	.004	.010	.005	.007	-.001	-.002	-.013	-.002	.095	-.004	-.009	-.002
Item 12	-.017	.004	.010	-.007	.011	.020	.005	.008	.023	.003	-.004	.126	.006	.232
Item 13	-.033	.009	.000	.012	.026	.004	-.002	.044	.014	.005	-.009	.006	.232	.015
Item 14	-.009	-.006	-.003	.003	.009	.006	.000	.007	.003	.000	-.002	.002	.015	.023

1999 & 2001 Covariance Matrix

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14
Item 1	-.004	-.018	-.001	.000	-.033	-.016	-.002	-.031	-.004	-.002	-.009	.007	-.020	.010
Item 2	.004	.039	-.001	.018	.031	.015	.003	.035	-.001	-.012	.031	-.002	-.008	.015
Item 3	-.007	-.012	.015	.004	-.006	-.006	-.001	.002	-.006	.007	-.014	.017	-.003	.003
Item 4	-.007	.012	-.009	.004	.011	.006	-.001	.025	.000	.001	.004	.017	.020	-.008
Item 5	.011	.026	.010	.012	.012	.014	-.001	.006	-.016	.003	.017	-.006	-.006	.010
Item 6	.015	-.015	.016	-.001	.004	.057	-.002	.029	.033	.011	.006	.015	.009	.001
Item 7	-.001	.003	-.001	.003	-.001	-.002	.000	.002	-.002	.004	.005	.005	.003	.000
Item 8	-.001	.026	-.012	.008	.027	.049	-.001	.027	.013	.033	.021	-.001	.024	.006
Item 9	.005	.002	.010	.018	.006	.002	-.001	.018	.054	.015	-.001	.011	.011	-.008
Item 10	-.002	-.002	-.004	.004	-.002	-.001	.000	.000	.001	.005	-.004	-.002	.004	.005
Item 11	.007	.016	.003	-.004	.009	.003	-.001	.006	-.017	-.008	.008	-.005	-.011	.017
Item 12	.015	.009	.034	.025	.015	.029	-.001	.014	.024	.006	.012	.015	.019	.008
Item 13	-.007	.028	-.003	.003	.031	.015	-.002	.051	.023	-.036	.004	.011	.072	.012
Item 14	-.002	-.005	.006	.008	.008	.009	.000	.008	.005	.003	.013	.004	.002	-.002

Appendix C

SYNTAX USED TO CALCULATE ASE AND CONFIDENCE INTERVALS[9]

```
matrix.
compute numbitem = 14.
compute numbsubj = 171.
compute itemcov = {.241, -.014, .001, .001, .004, .018, -.002, -.028, -.014, -.007, .028, -.017, -
.033, -.009; -.014, .251, .007, .025, .010, .005, .003, .025, -.002, .003, .007, .004, .009, -.006;
.001, .007, .113, .025, .008, .013, -.001, -.007, .002, .004, .004, .010, .000, -.003; .001, .025,
.025, .113, .008, .001, -.001, .005, -.010, -.002, .010, -.007, .012, .003; .004, .010, .008, .008,
.108, .026, -.001, .006, .002, .004, .005, .011, .026, .009; .018, .005, .013, .001, .026, .195, -.002,
.009, .015, -.005, .007, .020, .004, .006; -.002, .003, -.001, -.001, -.001, -.002, .006, -.001, -.001,
.000, -.001, .005, -.002, .000; -.028, .025, -.007, .005, .006, .009, -.001, .153, .018, .003, -.002,
.008, .044, .007; -.014, -.002, .002, -.010, .002, .015, -.001, .018, .108, .004, -.013, .023, .014,
.003; -.007, .003, .004, -.002, .004, -.005, .000, .003, .004, .017, -.002, .003, .005, .000; .028,
.007, .004, .010, .005, .007, -.001, -.002, -.013, -.002, .095, -.004, -.009, -.002; -.017, .004, .010, -
.007, .011, .020, .005, .008, .023, .003, -.004, .126, .006, .232; -.033, .009, .000, .012, .026, .004,
-.002, .044, .014, .005, -.009, .006, .232, .015; -.009, -.006, -.003, .003, .009, .006, .000, .007,
.003, .000, -.002, .002, .015, .023}.
compute one=make(numbitem, 1,1).
compute jtphij=transpos(one).
compute jtphij = jtphij * itemcov.
compute jtphij =  jtphij * one.
compute trmy=trace(itemcov).
compute trmy=trmy/jtphij.
compute myalpha=1-trmy.
compute nn1=numbitem-1.
compute nn1=numbitem/nn1.
compute myalpha=nn1 * myalpha.
compute trphisq=itemcov*itemcov.
compute trphisq=trace(trphisq).
compute trsqphi=trace(itemcov).
compute trsqphi=trsqphi**2.
compute ttp=itemcov * itemcov.
compute jtphisqj=transpos(one).
compute jtphisqj=jtphisqj * ttp.
compute jtphisqj=jtphisqj * one.
compute omega=trphisq+trsqphi.
compute omega=jtphij * omega.
compute omegab=trace(itemcov).
compute omegab=omegab * jtphisqj.
compute omega=omega-(2*omegab).
compute omega=(2/(jtphij**3))*omega.
compute s2=(numbitem**2) / ((numbitem-1)**2).
compute s2=s2*omega.
compute se=sqrt(s2/numbsubj).
compute cimin95=myalpha-(1.96*se).
compute cimax95=myalpha+(1.96*se).
print myalpha /format ="f8.3"/title= 'Your coefficient alpha is:'.
```

[9] Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods, 8,* 88-101. Hakstian, A. R., & Whalen, T. E. (1976). A K-sample significance test for independent alpha coefficients. *Psychometrika, 41,* 219-231.

```
print cimin95 /format = "f8.3"/title= 'The lower 95% confidence limit follows:'.
print cimax95 /format = "f8.3"/title= 'The upper 95% confidence limit follows:'.
end matrix.
```

Appendix D

Candidates' Race and Rank

Candidate ID	Race	Rank
7	White	1
13	White	1
35	White	1
42	White	1
59	White	1
72	White	1
100	White	1
103	White	1
108	African American	1
120	White	1
123	White	1
142	White	1
143	White	1
11	White	14
12	White	14
18	African American	14
20	White	14
22	White	14
30	White	14
36	White	14
38	White	14
49	White	14
53	White	14
66	African American	14
71	White	14
78	White	14
79	White	14
81	White	14
88	White	14
90	African American	14
93	White	14
104	White	14
109	White	14
110	White	14
113	White	14
114	White	14
118	African American	14
122	White	14
124	White	14
127	White	14
132	African American	14
133	White	14
134	White	14
135	African American	14
144	White	14
146	White	14
154	African American	14
159	White	14
162	White	14
167	White	14
168	African American	14
6	White	15
10	White	53
14	White	53
17	African American	53
19	White	53
23	African American	53
24	African American	53
25	White	53
26	African American	53
33	White	53
39	White	53
40	White	53
41	White	53
51	White	53
54	White	53
57	White	53
58	White	53
60	White	53
62	White	53
67	White	53
69	White	53
70	White	53
74	White	53
83	White	53
86	African American	53
89	African American	53
96	White	53
98	White	53

Candidate ID	Race	Rank
101	African American	53
105	White	53
107	African American	53
112	African American	53
126	African American	53
128	African American	53
131	White	53
137	White	53
139	White	53
147	African American	53
156	African American	53
160	African American	53
161	White	53
164	White	53
166	White	53
170	African American	53
3	African American	96
9	African American	96
15	White	96
16	White	96
21	African American	96
27	White	96
43	White	96
44	African American	96
45	African American	96
46	African American	96
48	White	96
52	African American	96
55	White	96
61	African American	96
68	White	96
76	African American	96
80	African American	96
84	African American	96
87	White	96
95	African American	96
106	White	96
116	White	96
117	White	96
121	White	96
125	White	96
136	African American	96
138	African American	96
145	White	96
148	African American	96
150	African American	96
158	African American	96
171	White	96
4	African American	128
8	White	128
29	White	128
31	White	128
32	White	128
37	White	128
47	African American	128
50	African American	128
63	African American	128
73	African American	128
75	White	128
77	White	128
82	African American	128
91	African American	128
94	White	128
99	African American	128
102	White	128
111	White	128
129	African American	128
140	African American	128

Candidate ID	Race	Rank
149	African American	128
163	African American	128
165	African American	128
2	African American	151
5	African American	151
34	African American	151
56	African American	151
64	African American	151
65	White	151
97	African American	151
115	White	151
119	White	151
141	African American	151
153	White	151
157	White	151
169	African American	163
1	White	163
28	African American	163
85	African American	163
92	African American	163
130	White	163
152	African American	163
155	African American	163
151	African American	171

Table 1

Racial differences using different selection techniques

	d score		
Selection Technique	**White-Black**	**White-Hispanic**	**Meta-Analysis**
Cognitive Ability	1.10	.72	Roth, BeVier, Switzer, & Tyler (2001)
GPA	.78	N/A	Roth & Bobko (2000)
Job Sample/Job Knowledge	.38	.00	Schmitt, Clause, & Pulakos (1999)
Biodata	.33	N/A	Bobko, Roth, & Potosky (1999)
Structure Interview	.23	N/A	Huffcut & Roth (1998)

d score represents the difference between standardized population means; Source: Aamodt (2004)

Table 2

Hypothetical Score Distribution and Test Score Use (from page X)

Test Score	**Race**	**Selection Possibility – Top Down**	**Selection Possibility – Banding**
95	Caucasian	X	X
94	Caucasian	X	X
94	Caucasian	X	X
92	Caucasian	X	X
91	Caucasian	X	X
89	African American		X
89	Caucasian		X
89	Caucasian		X
87	African American		X
87	African American		X
86	African American		X
86	Caucasian		X
86	African American		X
85	Caucasian		
85	African American		
83	Caucasian		
82	Caucasian		
82	African American		
81	Caucasian		
80	Caucasian		
80	African American		
80	Caucasian		
80	Caucasian		
79	African American		
79	Caucasian		
Percentages	**64% - Caucasian 36% - African American**	**100% - Caucasian 0% - African American**	**62% - Caucasian 38% - African American**

Table 5

KSA and Assigned Testing Modalities for Sergeant Selection Procedures

Knowledge/Skill/Ability	**Written (M.C.) Test**	**In-Basket/Work Sample**	**Oral Examination**
Technical Knowledge	√		
Oral Expression			√
Written Expression		√	
Interpersonal Relations			√
Information Analysis		√	√
Judgment & Decision Making		√	√
Planning & Organizing		√	√
Resource Management		√	√

Figure 1

Item covariance matrix for test-retest data (from Green, 2003)