Abstract

Corrections to Twenty Years of Banding:

The Necessity of Precision

by Sean M. Gasperson

April, 2010

Director: Mark C. Bowler, Ph.D.

DEPARTMENT OF PSYCHOLOGY

Statistically-based banding is often considered a viable method for minimizing adverse impact in test-based decisions. By utilizing the standard error of the difference (SED), scores are equated based on the assumption that there is substantial unreliability in any single observed score. However, as noted by Dudek (1979), the current procedure for calculating the standard error of measurement (SEM), which is used when calculating SED, is erroneous. Concurrently, the tendency of score to regress to the mean with multiple administrations of a test is a problem that is often overlooked in literature (Smith & Smith, 2005). This study compares the differences between banding based selection decisions when these corrections are applied. In general, results suggest that applying these corrections to methods of calculating SED can produce variations in the decisions. These potential legal and ethical implications of these discrepancies are discussed.

Corrections to Twenty Years of Banding:

The Necessity of Precision

A Thesis

Presented To

The Faculty of the Department of Psychology

East Carolina University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts in Psychology

by

Sean M. Gasperson

April, 2010

CORRECTIONS TO TWENTY YEARS OF BANDING:

THE NECESSITY OF PRECISION

by

Sean M. Gasperson

APPROVED BY:

DIRECTOR OF DISSERTATION:_____

Mark C. Bowler, Ph.D.

COMMITTEE MEMBER:_____

Karl L. Wuensch, Ph.D.

COMMITTEE MEMBER:_____

Jennifer L. Bowler, Ph.D.

CHAIR OF THE DEPARTMENT OF PSYCHOLOGY:

_____

Kathleen A. Row, Ph.D.

DEAN OF THE GRADUATE SCHOOL:

_____

Paul J. Gemperline, Ph.D.

Acknowledgements

I would like to thank Dr. Mark Bowler for all of his guidance and structure while directing this thesis. I would also like to thank my committee members, Drs. Jennifer Bowler and Karl Wuensch, for their insight that they contributed to this paper. Special thanks to Dr. Wuensch, for keeping statistics both relevant and interesting.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I: INTRODUCTION

Test score banding, the clustering similar test scores into groups, is a concept that is common in modern society (Murphy & Myors, 1995). Within the realm of employment decisions, statistically-based banding has been utilized over the past twenty years to serve as an alternative to strict top-down selection (Cascio, Outtz, Zedeck, & Goldstein, 1991; Henle, 2004). More specifically, it is commonly used as a method for deploying affirmative action programs and as by organizations under a consent decree (Barrett & Louke, 2004). However, the use of statistically-based banding has not gone without debate in both the research literature (Cascio, Goldstein, Outtz, & Zedeck, 1995) and the legal system (Henle, 2004). Critics of banding cite the inappropriateness of the procedure, as well as a perceived contradiction between the rationale of banding and the method with which it is employed (Schmidt, 1991; Schmidt & Hunter, 1995). Regardless of this debate, one primary issue that has not been addressed is the formula for the standard error of measurement (SEM) that is to be used for calculating bandwidth (Cascio et al., 1991).

With statistically-based banding, the fundamental purpose is to generate a range of possible true scores around an observed score so that scores within the band can be equated. However, based on the derivations of Dudek (1979), the formula that is commonly used to calculate the SEM in banding is inappropriate. That is, the formula that is commonly used to calculate SEM is for generating a range of observed scores around a true score. In fact, a formula for generating a range of true scores around an observed score is more appropriate. An additional psychometric issue relating to banding is the tendency of scores to regress to the mean (Rocconi & Ethington, 2009; Smith & Smith, 2005). That is, when individuals complete a measure multiple times, the observed scores should most likely shift towards the mean score.

Given the legal concerns surrounding employment decisions, as well as the gravity of the decision of who will or will not be selected, the correct SEM formula, as well as accounting for regression to the mean (RTM), should be employed when using statistically-based banding. The present study addresses both of these issues. First, the principles of statistical-based banding are discussed along with the traditional formulas that are used. Second, we detail the debate regarding the appropriateness of the use of banding. We then turn to the different formulations of the SEM – as well as the calculations of RTM – and discuss their impact on the process of banding. Finally, these methods are applied to two unique samples in an attempt to evaluate the differences that they engender in employment decisions.

Banding

Banding, in its traditional form, is evident in many facets. For example, a preset range of scores on a particular test could be labeled as "pass/fail," or broken down to A, B, C, D, and F (Henle, 2004). In the later case, all scores on a test falling within a particular range will be assigned one of five letter grades. By this standard, the highest score in the range receives the same grade as the lowest score in the range (e.g., both a 100% and a 90% receive an "A"). However, as noted by Bobko and Roth (2004), these groupings "are an administrative convenience only; they are not necessarily justified on statistical and psychometric bases" (p. 292). Yet, they commonly are accepted without reservation. Less commonly accepted – and often debated – is statistically-based banding.

Statistically-based banding was developed as an alternative to top-down referral and is intended to reduce adverse impact while providing a statistical justification for any loss of utility (Bobko, Roth, & Nicewander, 2005; Cascio et al., 1991). As noted by Cascio et al. (1991), any referral method other than top-down selection will result in the loss of utility (i.e., poorer job

performance from those selected). Regardless, banding is considered by most to be a viable method by which organizations can attempt to maximize diversity while minimizing adverse impact. However, the key questions that banding research seeks to answer are (1) to what extent can loss of utility be minimized while still effectively reducing adverse impact and (2) what is the most psychometrically appropriate method for determining the width of the band?

Psychometrics of Banding

The procedure of banding operates under two main assumptions. The first is that in every measurement, there is a certain degree of error. That is, a measure is never completely accurate is its measurement of an individual. Second, an individual's scores on a measure – when given the measure multiple times – will vary. Thus, the degree of error in an individual's test scores is measured by the standard error of measurement (SEM) which is a function of the reliability of the measure (r_{xx}) and the standard deviation (SD_x) of obtained scores (Ghiselli, Campbell, & Zedeck, 1981). The traditional formula for the standard error of measurement (SEM) is:

$$\text{SEM} = SD_x \sqrt{1 - r_{xx}} \qquad (1)$$

As noted by Cascio et al. (1991), the standard error of measurement is "an estimate of the standard deviation of the normal distribution of test scores that an individual would obtain if he or she took the test an infinite number of times" (p. 240). Therefore, accepting the appropriateness of this formula, in a normal distribution, it may be assumed that an individual's observed score will only infrequently (5% of the time) deviate more than 1.96 SEMs from any given observed score.

Given that an individual's true score is a function of his or her observed score plus error, a key question remains: Can it be assumed that differences between two individuals' observed scores will be an indication of actual differences on a criterion? When a measure has complete

criterion-related validity this will be the case (consistent with the general linear model); however, a measure's validity will almost certainly fall below one (Cascio et al., 1991). Therefore, Gulliksen (1950) developed the standard error of the difference (SED) between two scores, which establishes a confidence interval for the difference between the true scores of two individuals. The complete formula for SED is

$$X_i - X_j + C \times \text{SEM}\sqrt{2} > T_i - T_j > X_i - X_j - C \times \text{SEM}\sqrt{2} \qquad (2)$$

where C represents the point chosen on a z distribution corresponding to a desired confidence level. There is substantial debate regarding the appropriate confidence level to choose (see Cascio & Zedeck, 1983). Typically, a 95% confidence interval is chosen. The confidence interval must be chosen based on whether Type I errors (erroneous acceptances) or Type II errors (erroneous rejections) are more serious. A higher alpha will decrease the size of the band, and a lower alpha will increase the size of the band (Cascio & Zedeck, 1983). Furthermore, if this confidence interval includes zero, it cannot be said with a level of confidence (corresponding to C) that the true scores of individual i and individual j are significantly different.

Banding uses the aspect of the SED formula $C \times \text{SEM}\sqrt{2}$ to determine a range of observed scores that cannot be considered significantly different due to the possible range of true scores (Cascio et al., 1991). Scores falling outside the range $C \times \text{SEM}\sqrt{2}$ can be considered significantly different from the referent score. Thus, due to measurement error and the lack of complete reliability in any measure, it is possible that the true score of individual i may be higher than the true score of individual j, even if the observed score of individual i is lower than that of individual j, as long as the two observed scores fall in the range $C \times \text{SEM}\sqrt{2}$.

Use of Banding

As previously noted, the use of strict top-down selection often results in adverse impact, and any method of selection other than strict top-down will result in a loss in utility (Cascio et al., 1991). Statistically-based banding offers an alternative method of selection that seeks to reduce adverse impact while minimizing the loss of utility. Cascio, et al. (1991) evaluated two types of statistically-based banding: fixed bands and sliding bands. With the *fixed banding* procedure, a band of scores is calculated using the SED formula. Using the highest score as a starting point, selections are then made from within the band until all of the positions have been filled or the band is exhausted. If the band is exhausted before all of the positions are filled, a second band of scores is calculated using the top score immediately outside the first band as the referent score. This process continues until all of the positions are filled or there are no remaining applicants.

The second type of statistically-based banding is *sliding bands*. This procedure differs from the fixed banding procedure in that after the top score from within the band is selected, the band shifts (i.e., "slides") down and uses the next highest score as the new starting point for the band. Thus, scores that were immediately outside the original band are included in the new band. Scores are selected from the new band until the top score in the band is selected, in which case the band slides down again to include the next score immediately outside the band (Cascio et al., 1991). This process continues until all of the positions are filled or there are no remaining applicants.

Selection methods from within bands have been proposed to be either random or based on secondary criteria (Cascio et al., 1991). According to Barrett and Lueke (2004), some methods of secondary selection criteria "are specific to the city or organization and involve additional factors such as veteran status or seniority when making promotion decisions" (p. 97).

Minority status may also be a contributing factor in secondary criteria, although it may not legally be the sole source of selection within a band if the organization is not outwardly engaged in an affirmative action program or under a consent decree (Henle, 2004).

The Banding Debate

Over the past twenty years, the use of banding in employee selection has been a highly debated topic (cf., Cascio et al., 1991; Murphy & Myors, 1995; Schmidt, 1991; Schmidt & Hunter, 1995). Cascio et al. (1991) introduced the concept of sliding bands as an alternative to top-down selection and within group norming in order to reduce adverse impact and minimize the loss in utility. They purported that sliding bands provided significant advantages over top-down selection and fixed banding, in that greater number of protected groups would be incorporated into the workforce at a faster rate, and adverse impact may be eliminated from a selection procedure that typically produces adverse impact. When comparing the different banding methods, Cascio et al. (1991) concluded that "the use of the sliding-band procedure reconciles economic and social objectives within the framework of generally accepted procedures for testing hypotheses about differences in individual test scores" (p. 263). This conclusion was supported by Sackett and Roth (1991), whose Monte Carlo comparisons of many selection procedures concluded that sliding bands produced more minority group hiring than any other banding approach.

Schmidt (1991) responded to Cascio et al. (1991)'s sliding band procedure with three primary concerns. The initial critique cited the irrelevance of statistical significance testing in employee selection. According to Schmidt, the basic selection model is based on a regression equation predicting performance. Therefore, it is inherently based on probabilities. The determination of significant differences between predictor scores is irrelevant, because in any

test with appropriate criterion-related validity, the individual scoring higher should – over time – outperform an individual scoring lower on a predictor. Zedeck, Outtz, Cascio, and Goldstein (1991) provided a response to this argument noting that top-down selection is the most optimal method of selection and that the regression model will over time likely be accurate for a group of applicants. However, organizations could have other goals in mind during selection, such as workforce diversity.

The second primary criticism attacks the logic of banding. Schmidt (1991) stated that the use of banding inherently contradicts its own underlying theory, in that the highest score in a band is not considered significantly different from the lowest score in the band, yet the lowest score in the band is considered significantly different from the score immediately outside the band. If the logic of banding were applied appropriately, Schmidt states, the score immediately outside the band would be included in the band because it does not lie outside the range of significance of most of the scores in the band. This process would be applied with the next score immediately outside the band, and so on. Eventually, all scores would be included in the band, leading to random selection. Murphy and Myors (1995) labeled Schmidt's argument a "slippery slope" argument in that the basis that interpreting the SED argument to include the complete range of scores in the band presents a logical fallacy as there is an inherent problem in deeming the top score equal to the bottom score of a distribution. Banding is, according to Murphy and Myors, inconsistent in the aspect that the SED is not applied to all scores in the distribution. However, they note that there are few selection systems or decision models that attain the status of complete consistency, and to deem banding's inconsistency as a fatal logical flaw would be inappropriate. Additionally, Murphy and Myors (1995, p. 196) cite current widely used decision-making systems that "fail to meet the consistency criterion," such as academic grading scales

(grades 90 and 100 are both given the same letter grade, while 89 and 90 are given different letter grades) and multiple choice tests. Although Murphy and Myors concede that banding is not without problems, the critique offered by Schmidt should not be considered a fatal flaw to the procedure.

Finally, Schmidt (1991) refuted the data that was used by Cascio et al. (1991) in that it appeared to be anomalous. Specifically, he noted that in most data sets, minorities typically have a smaller standard deviation than those in the majority. However, in the data used by Cascio et al. (1991), the standard deviation of the minority groups were all higher than the standard deviation of the nonminority (i.e., White) group (i.e., the average s for minority groups was 13.85 and the s for the nonminority group was 10.40). Schmidt (1991) contended that the conclusions drawn by Cascio et al. (1991) regarding the effectiveness of sliding bands were inappropriate due to the atypical data. In response to Schmidt's contention of the data being aberrant, Zedeck et al. (1991) offered examples of data from varying samples with similar proportions of standard deviations. In addition, it was presented that recent literature has suggested conflicting conclusions. Therefore, to deem the data used by Cascio et al. (1991) anomalous would be erroneous.

Schmidt and Hunter (1995) provided a rebuttal to the numerous criticisms surrounding Schmidt's (1991) initial critique of banding, stating that the argument was misunderstood. According to Schmidt and Hunter, the fatal flaw in banding comes not from its statistical rationale, but from the contradiction between its statistical rationale and its operational procedures. The statistical rationale in banding is based on the SED, which determines a range of scores that are not statistically different from one another. However, Schmidt and Hunter argue that operationally, there is a distinction made between the lowest score in the band and the score

immediately outside the band, even though the two scores are not statistically different according to the SED. Although Murphy and Myors (1995) contend that this inconsistency is not uncommon and therefore not a fatal flaw, Schmidt and Hunter claimed that other banding systems to not have the same statistical rationale as statistically-based banding (e.g., groups/bands are predetermined, and not justified by providing evidence of the similarity between two scores).

In an attempt to provide further clarification to the issues surrounding banding, Cascio, Goldstein, Outtz, and Zedeck (1995) responded to the debate surrounding banding. Specifically, Cascio et al. (1995) emphasized the fact that Schmidt misconstrued the operational procedures of banding, in that the top score in the band is the referent score, rather than any two scores within a distribution. Concurrently, Cascio et al. (1995) reiterated that there is a well-defined set of procedures when using sliding bands, and critics often overlook this fact. A clear distinction was also made between statistically-based banding and a bonus point system. Rather than arbitrarily giving bonus points to a particular group (i.e., minorities), banding utilizes statistical methodology to establish a range of scores whose criterion could vary, and allows for consideration of other factors in the selection decision.

In an attempt to provide a review of the banding debate, Campion et al. (2001) published an article in a question and answer format. By posing ten questions dealing with the controversial topic, the proponents of banding (e.g., Outtz & Zedeck), and the critics of banding (e.g., Schmidt), and neutral observers (e.g., Murphy) were able to clarify their positions. These questions addressed topics such as the psychometric pros and cons of banding, recommendations regarding the use of banding, potential alternatives to banding, and the legality of banding. As would be expected, proponents of banding stated the psychometric appropriateness of banding,

and critics refuted it. However, there appeared to be a few general points of agreement among all experts. These included the fact that there should be multiple considerations in employment decisions, that score aggregation is common, and that deviating from the general linear model would lead to a loss of utility.

Banding Utility Losses

Siskin (1995) provided a mathematical model that computed the expected difference in performance between the top and bottom ranked persons in the band, as well as the probability that the top-ranked individual will outperform the lowest ranking person in the band under various situations. Although it is generally accepted that the highest scoring person in the band will typically outperform the lowest scoring person in the band, Siskin sought to discover how much of a difference there would actually be.

To do so, Siskin (1995) constructed three mathematical models. These models (a) define true overall performance, consisting of the criterion measure and the remaining unmeasured performance, (b) define the actual test score, which consists of a "true" test score plus random error, and (c) "expresses the criterion measure as a linear function of the true test score plus error" (Siskin, 1995, p. 216). The three models are as follows:

$$P_i = C_i + \delta_i \tag{3}$$

$$T_i = \tau_i + \Sigma_i \tag{4}$$

$$C_i = \alpha + \beta\tau_i + \Delta_i \tag{5}$$

where

P_i = true job performance

C_i = measured component of job performance

δ_i = unmeasured component of job performance

T_i = observed test score

τ_i = true test score

Σ_i = error in test measurement

Δ_i = error in prediction

Siskin (1995) further notes:

$$\delta_i \sim N(0, \sigma_{\delta_i}^2) \tag{6}$$

$$\Sigma_i \sim N(0, \sigma_{\Sigma_i}^2) \tag{7}$$

$$\Delta_i \sim N(0, \sigma_{\Delta_i}^2) \tag{8}$$

δ_i, Σ_i, and Δ_i are independent (p. 216).

The following equation provided by Siskin provides a method for calculating the predicted likelihood that the top scorer in the band will perform better than the person d units of SED below that scorer:

$$\Pr\left[(P_i - P_j) > 0 \big| (T_i - T_j)\right] = Z\left[(-d\sqrt{(p_a^2 p_b^2 - p_a^2 p_b^2 p_c^2))} \div (\sqrt{(1 - p_a^2 p_b^2 p_c^2))}\right] \tag{9}$$

where

p_a^2 = the correlation squared (R^2) between the measured job component and true

job performance

p_b^2 = the correlation squared (R^2) between true test score and measured

performance

p_c^2 = the correlation squared (R^2) between true test score and observed test score

Z = cumulative standard normal variable (Siskin, 1995, p. 217).

Finally, he provided the following equation that calculates "…the expected difference in performance between the top-ranked scorer in the band and the person d units of SED below

him, assuming the reliability of the test, the true validity of the test to predict the performance criteria, and the extent to which the criteria measure overall performance (criterion sufficiency)" (p. 217):

$$E\left[P_i - P_j \mid T_i - T_j\right] = \left(-d\sqrt{2(1 - \rho_c^2)\rho_b^2\rho_a^2}\right)(\sigma_p) \tag{10}$$

By applying these formulas to varying combinations of band widths (in units of SED), test reliability coefficients, criterion sufficiency, and validity coefficients, Siskin's (1995) results suggested that the difference in expected units of performance between the top scorer in the band and the lowest scorer were often quite small (usually between .1 and .25 standard deviations). Furthermore, the likelihood that the top scorer would outperform the lowest scoring person averaged between .6 and .7. Siskin's analysis led him to conclude that "Given the value of the social objectives that can be accomplished with banding, and the small expected loss in performance, banding seems to offer a logical, fair, and reasonable approach to the trade-off between social and economic objectives" (Siskin, 1995, p. 226). His results "…support the use of banding and selection based on secondary criteria and generally show that the social gains of banding may be greater than the economic cost" (Siskin, 1995, p. 215).

Legal Status of Banding

After within-group percentile norming was deemed illegal by the Civil Rights Act of 1991, banding was proposed as an alternative selection method to top-down referral (Cascio et al., 1991). Henle (2004) reviewed the legal status of banding. An important question that was addressed was whether or not banding complied with the Civil Rights Act of 1991. Section 106 of the Civil Rights Act, addressing score adjustment, states:

> It shall be an unlawful employment practice for a respondent, in connection with
>
> the selection or referral of applicants or candidates for employment or promotion,

to adjust the scores of, use different cutoff scores for, or otherwise alter the results

of, employment related tests on the basis of race, color, religion, sex, or national

origin.

Furthermore, Section 107 (a) states:

Except as otherwise provided in this title, an unlawful employment practice is

established when the complaining party demonstrates that race, color, religion,

sex, or national origin was a motivating factor for any employment practice, even

though other factors also motivated the practice.

After reviewing several case studies, Henle summarized that "banding has not been found to

violate Section 106 or 107 (a) of the Civil Rights Act of 1991. In fact, the Seventh and Ninth

Circuit have upheld its legality under this Act" (2004, p. 424).

Henle (2004) also examined another issue – minority preference within bands. It would

appear that banding violates the principle of score adjustment based on minority preference.

However, the legality of banding remains intact as long as race, sex, etc. is not the *sole* criterion

of selection from within the band (Hartigan & Wigdor, 1989; Henle, 2004). Henle also addresses

the relationship between the legality of banding and affirmative action, stating that "the legality

of banding has been upheld regardless of whether it is used in conjunction with consent decrees,

court orders, or voluntary affirmative action plans if selection from within the band is not based

exclusively on minority preference. However, past discrimination will not shield organizations

from legal challenge when they use only minority preference to select from within bands" (2004,

p. 426).

Much of the literature regarding the legality of banding focuses on three specific court

cases (Barrett & Louke, 2004). In *Chicago Firefighters v. City of Chicago* (1999), a group of

white firefighters claimed that their equal-protection rights were being infringed upon by the affirmative-action promotions of minority firefighters. The appeals court upheld the ruling of the district court, ruling that banding was not identical to race norming, nor was it a form of score adjustment. In *Bridgeport Guardians, Inc. v. City of Bridgeport* (1991), minority police officers contested that a promotional examination did not distinguish between qualified and unqualified applicants. The courts ruled that there was adverse impact in the selection system, and that banding served as an effective alternative selection procedure. Courts also deemed banding as acceptable in *Officers for Justice v. Civil Service Commission* (1992), when it was proposed as a method of promoting a higher percentage of minorities.

Although these three cases are the main foci of the literature, a search of legal cases pertaining to banding conducted by Barrett and Louke (2004) found over 30 court cases that mentioned banding. These cases refer to banding in a variety of ways, ranging from being proposed as an alternative selection procedure to being challenged by a group (nonminority and minority alike) questioning the fairness of the procedure. Overall, there does not appear to be a consensus regarding the legality of banding. Most recently, the opinion of the court in *Ricci v. DeStefano* (2009) stated that although banding had been disallowed by a local legal statute in that particular case, it could be a viable alternative selection model.

Corrections to Banding and Use of Precision

Standard error of measurement. As previously noted, statistically-based banding utilizes the standard error of measurement (SEM) formula to calculate bandwidth (Cascio et al., 1991). The SEM formula used by Cascio et al. (1991) is SEM = $SD_x\sqrt{1 - r_{xx}}$. However, Dudek (1979) proves this interpretation of SEM to be erroneous. The SEM formula used by Cascio et al. (1991) provides a band of possible *observed scores around a true score*. However, the basic assumption

of banding is that the given score of any individual is inaccurate (i.e., it is an observed score and not a true score). From this assumption, the nature of banding is to develop a range of possible *true scores around an observed score*. Thus, calculating a band of observed scores around a true score is a psychometrically inappropriate because if you have the true score then banding is unnecessary. Thus, the appropriate interpretation of the SEM would be to calculate a band of true scores about an observed score that is held constant. This is estimated by the following formula:

$$\text{SEM} = SD_x\sqrt{r_{xx}(1 - r_{xx})}, \tag{11}$$

This more accurately assesses an individual's possible range of true scores around an observed score. When applied to the calculation of the SED the inclusion of the additional reliability term makes the bands smaller. Thus, there may be substantial difference when utilizing this more appropriate formulation of the SEM.

Regression to the mean. Generally speaking, regression to the mean is defined as the tendency for an individual's test score to move closer to, or "regress," towards the mean of the score distribution after multiple administrations of the test (Rocconi & Ethington, 2009). Regression to the mean has been seen as a threat to internal validity and is often overlooked in literature (Smith & Smith, 2005). The foundation of the regression to the mean lies in the relationship between an individual's true score and observed score (Krause, 2008). The formula for the true score of examinee *j* is defined as

$$T_j = X_j + E_j, \tag{12}$$

where T_j = the true score of examinee *j*, X_j = the observed score of examinee *j*, and E_j = the error present in the observed score of examinee *j* (Crocker & Algina, 2008). In accordance with modern test theory, the error component of an individual's test score is directly affected by the measure's reliability. Tests with higher reliability will likely result in smaller error, and tests with

lower reliability will likely result in larger error. As no measure has perfect reliability, and the error component will likely never be reduced to zero, an individual's test score will likely vary with multiple administrations, most often shifting towards the mean (Rocconi & Ethington, 2009).

In reviewing current literature, the concept of regression to the mean is statistically supported, yet highly overlooked (Smith & Smith, 2005). The formula for the adjustment of test scores for regression to the mean, as outlined by Roberts (1980), is

$$x' = x + (1 - r_{xx})(\mu - x), \tag{12}$$

where x' = the score regressed to the mean, x = the observed test score, r_{xx} = the reliability of the test measure, and μ = the mean of the sample tested. When applied to the calculation of the SED the inclusion of RTM makes the bands shift. Thus, there may be substantial difference when utilizing this more appropriate formulation of the SEM.

Overview of Purpose

It is necessary to modify current banding procedures in such a way as to be as precise and accurate as possible when utilizing statistically-based banding as a selection system for the following reasons: (a) It is evident that the methods and use of banding have been highly controversial in literature (Cascio et al., 1991; Schmidt, 1991; Schmidt & Hunter, 1995) as well as in the legal system (e.g., *Chicago Firefighters v. City of Chicago,* 1999; *Officers for Justice v. Civil Service Commission,*1992); (b) Using an inappropriate method to determine bandwidth (SEM) as well as arbitrary rounding (e.g., 1.96 = 2) opens the door to a tremendous amount of legal scrutiny; (c) The gravity of the situation is great, given that there will often be applicants who are rejected from a job or promotion; and (d) The Ethical Principles of Psychologists and Code of Conduct hold psychologists to such, stating that "Psychologists seek to promote

accuracy, honesty, and truthfulness in the science, teaching, and practice of psychology" (APA, Principle C, italics added). The purpose of this study is to demonstrate the effect that proper application of statistically-based banding (i.e., including RTM and the correction formulation of SEM) has on banding-based selection decisions, compared to current procedures and other forms of selection methods.

CHAPTER II: METHOD

Data

Two separate sets of data were analyzed. The first data set was taken from Cascio et al. (1991). These data represent scores on a selection measure for 3,377 candidates for positions as firefighters in a large U.S. city. Table 1 provides descriptive statistics of the test scores for the six different ethnic groups in this data set. Thus, the initial set of analyses were conducted on the original data used to demonstrate the application of statistical banding in employee selection.

As previously noted, one of the criticism's of Cascio et al. (1991)'s application of statistically-based banding was that the data used may have been anomalous, in that the standard deviations of minority scores were larger than the standard deviation of nonminority scores (Schmidt, 1991). Specifically, Schmidt asserted that it is often the opposite that is true – that minority test scores typically have less variance than nonminority test scores. However, this assertion did not go unchallenged. Sackett and Roth (1991) suggested that minority and nonminority standard deviations are nearly equal and Zedeck et al. (1991) suggested that there is actually greater variance in minority test scores. Thus, is an attempt to utilize data that was representative of the current state of majority and minority test scores, a second data set was generated based on the results of Buttigieg (2006). Buttigieg presented the means and standard deviations of scores on the Wonderlic Personnel Test for 1,277 entry-level applicants at a manufacturing organization. Table 4 provides descriptive statistics of these test scores for multiple races. These means and standard deviations were engendered on the Cascio et al. (1991) data to create a second data set. However, as the Buttiegieg sample did not provide statistics for Filipinos, the target selection for Filipino and Asian applicants were added together. Furthermore, the raw scores from the Wonderlic data were doubled in order to set them to the

same scale as the Cascio et al. (1991) data and allow for easy comparison between the two sets of data.

Analyses

Before any alternate methods of calculating the SED were explored, a replication of the Cascio et al. (1991) analyses using the standard 95% confidence interval bandwidth was conducted. This set of analyses included strict top-down selection, within-group percentile, fixed bands (diversity-based), fixed bands (random), sliding bands (diversity-based), and sliding bands (random). The within-group percentile method of selection was not conducted as – under the Civil Rights Act of 1991 – this would be deemed a form of within group norming.

The second set of analyses involved the application of the corrected SEM formula – as proposed by Dudek (1979) – for calculating the SED to the original Cascio et al. (1991) data. As with the previous analyses, this set of analyses included strict top-down selection, within-group percentile, fixed bands (diversity-based), fixed bands (random), sliding bands (diversity-based), and sliding bands (random). For the third set of analyses, the test scores were first regressed to the mean using the formula presented by Rocconi & Ethington (2009). The numerous selection procedures were then applied to this data. The final procedure combined the appropriate SEM formula with the scores regressed to the mean. Thus, overall, four different methods of calculating bandwidth (i.e., original, new SEM, RTM, and new SEM and RTM) were compared using five different procedures of selection (i.e., strict top-down selection, diversity-based fixed bands, random fixed bands, diversity-based sliding bands, and random sliding bands).

Within band selection was concurrent with the procedures of Cascio et al. (1991). Each race had a target number of selectees that corresponded with the ratio of number of applicants from that race to the total number of applicants. Based on the labor market, the overall target

minority selection ratio was 47%. As in Cascio et al. (1991), selections were made for both 90

selectees and 216 selectees. However, there was one methodological difference from Cascio et

al. (1991). When the target selection ratio could not be satisfied for a specific race, Cascio et al.

(1991) simply selected the highest remaining minority in the band until the overall minority

selection ratio was satisfied (47%). This procedure is inappropriate – and would most likely not

be legally defensible – as it only serves the purpose of fulfilling the overall minority selection

ratio. Thus, if the target number of a particular race was not met, the highest remaining score in

the band was selected, regardless of race.

In order to compare each SED method and selection procedure, a number of statistics

were calculated. The first is the mean of the scores in the selection decision. This was done

simply by calculating the mean of the scores that were selected. The second is the lowest

percentile selected. Before any analyses were done, the overall percentile was calculated for each

raw score. The lowest percentile selected presents the overall group percentile that corresponds

with the lowest raw score in the selection decision. Additionally, the number of each race

selected as a result of each procedure was presented. The sum of all minorities selected within a

given procedure divided by the total number of selectees produced the minority selection ratio.

Based on the organization and community from which this data was originally drawn, the target

minority selection ratio was 47%.

In order to determine which minority selection ratios were significantly different from the

target selection ratio, 95% confidence intervals for the estimate of the population proportion

were calculated for each minority selection ratio. The formula used is as follows:

$$\hat{p} \pm Z \sqrt{\frac{p(1-p)}{n}} \qquad (13)$$

where $\hat{p}$ = the estimate of the population proportion, Z = the number of standard deviations required to include the desired percent confidence (in this case, 1.96), p = the proportion in the sample, and n = the sample size. Furthermore, correction for continuity was applied by adding $\frac{.5}{n}$ to the upper limit and subtracting $\frac{.5}{n}$ from the lower limit.

Raw scores were regressed to the mean to determine the effect that it has on the selection decision. This was done in accordance with Formula 12. Because selections by race are being compared, it would be most accurate to regress each score to the corresponding group mean (i.e., race mean). However, there are several issues that make this an unpractical alternative. First, this would have a stronger negative impact on minority members than majority members as the majority members would be regressed to a higher mean than minority members. Second, this type of group score alteration could be seen as a form of race norming. Thus, each raw score was regressed to the overall mean.

CHAPTER III: RESULTS

Cascio et al. (1991) Data

Table 1 provides descriptive statistics of the sample used by Cascio et al. (1991). It

should be noted that these data reflect Schmidt's (1991) criticism – that the standard deviation of

majority (White) applicants ($s = 10.40$) is lower than minority applicants (weighted $s = 13.78$).

Furthermore, the mean of majority applicants ($M = 75.60$) was 12.35 points higher than minority

applicants (weighted $M = 63.25$). These differences skew the distribution of data such that the

scores of White applicants are more heavily concentrated at the top of the distribution.

Table 2 provides the means, lowest percentile (the lowest total group percentile selected),

selection by race, minority selection ratio, and the 95% confidence intervals of the minority

selection ratios for each of the four SED methods, using the Cascio et al. (1991) data, selecting

for 90 employees. When comparing the new SEM method to the original method, the new SEM

method produced a higher selection mean ($M = 88.31$) than the original method ($M = 86.91$)

when employing the fixed bands (random) procedure. Similarly, for the diversity-based fixed

banding procedure, the new SEM method produced a higher selection mean ($M = 90.92$) than the

original procedure ($M = 90.66$). The new SEM method also produced a higher selection mean ($M

= 88.70$) than the original method ($M = 87.24$) when employing the sliding bands (random)

procedure. Mean differences did not exist between the new SEM and original methods for the

sliding bands (diversity-based) procedure. Although significance levels for mean differences

were not tested, this difference clearly indicate that employing the correct SEM formula yields a

different selection decision than the SEM formula used by Cascio et al. (1991).

Comparing regression to the mean procedures with original procedures, each procedure

employing RTM produced lower selection means than procedures employing the original

method. The fixed bands (random) procedure employing RTM yielded a lower selection mean ($M = 85.66$) than the original method ($M = 86.91$). Similarly, the diversity-based fixed banding procedure employing RTM also yielded a lower selection mean ($M = 89.61$) than the same procedure employing the original SED method ($M = 90.66$). The RTM sliding bands (random) procedure ($M = 85.80$) and RTM diversity-based sliding bands procedure ($M = 89.61$) yielded lower selection means than the original sliding bands (random) procedure ($M = 87.24$) and the original diversity-based sliding bands procedure ($M = 90.66$).

The new SEM plus RTM procedures were then compared to the original procedures. For the fixed bands (random) procedure, the new SEM plus RTM produced a higher selection mean ($M = 87.10$) than the original method ($M = 86.91$). However, for the diversity-based fixed banding procedure, the original SED method yielded a higher selection mean ($M = 90.66$) than the new SEM plus RTM method ($M = 89.61$). The original SED method also produced higher selection means for the sliding bands (random) procedure ($M = 87.24$) and the diversity-based sliding bands procedure ($M = 90.66$) than the new SEM plus RTM method employing the sliding bands (random) procedure ($M = 87.18$) and the diversity-based sliding bands procedure ($M = 89.61$).

The second criterion for comparison was the number of each race selected. As previously noted, the target number of selection for each race corresponded with the proportion of applicants for that particular race group. Comparing the new SEM method to the original method, the new SEM method selected 38 minorities while the original SEM formula selected 42 minorities when employing the diversity-based fixed banding procedure. This suggests that employing the uncorrected SEM formula resulted in the erroneous selection of four minorities and the erroneous rejection of four majority applicants. For the sliding bands (random)

procedure, employing the new SEM formula resulted in the selection of 17 minorities, compared to the 20 minorities selected using the original formula.

Comparing the RTM method with the original method, noticeable differences existed between the methods when employing the sliding bands (random) procedure. The RTM method resulted in five fewer minority selections. Furthermore, the RTM method yielded four Hispanic selectees whereas the original method yielded 13 Hispanic selectees. Differences in selections by race did not exist for diversity-based procedures. The new SEM and RTM method differed from the original method when employing the fixed bands (random) and the sliding bands (random) procedures, but did not differ when employing diversity-based procedures.

The minority selection ratio and 95% confidence intervals for each procedure are also noted in Table 2, as well as a graphic display of the intervals. This section of Table 2 is helpful in determining the relationship between the target minority selection ratio (the dotted line) and the actual minority selection ratio, as well as which proportions are significantly different. Although differences fell short of significance, the diversity-based fixed banding procedure employing the new SEM formula produced a lower minority selection ratio (42%) than the diversity-based fixed banding procedure employing the SEM method proposed by Cascio et al. (47%). As shown in Table 2, for every SED method, diversity-based fixed bands and diversity-based sliding bands provided significantly greater minority selection ratios than other selection procedures. Furthermore, in each SED method, the target minority selection ratio was either achieved or included in the 95% confidence interval of diversity-based fixed bands and diversity-based sliding bands.

Table 3 presents the means, lowest percentiles, selection by race, minority selection ratio, and 95% confidence intervals of the minority selection ratios of the four SED methods, using the

Cascio et al. (1991) data, selecting for 216 employees. Comparing the new SEM method with the original method, the new SEM method produced a higher selection mean ($M = 88.54$) than the original method ($M = 87.07$) when using the fixed bands (random) procedure. The new SEM method also yielded a higher selection mean ($M = 89.70$) than the original method ($M = 89.35$) when employing the diversity-based fixed banding procedure, as well as the sliding bands (random) procedure. Each procedure employing only RTM yielded lower selection means than procedures employing the original SED method. The same occurred for procedures employing the new SEM plus RTM, with the exception of sliding bands (random), in which the new SEM yielded a higher mean ($M = 86.70$) than the original method ($M = 86.55$).

These differences in selection decisions may be further explored by observing how many applicants were selected with each method by race. Comparing the new SEM method with the original method employing the corrected SEM formula resulted in 23 more majority selectees than the original method when using diversity-based fixed banding. The new SEM method also produces five more majority selectees when employing the fixed bands (random) procedure, as well as two fewer majority selectees when employing the sliding band (random) procedure. The RTM method also produces differences from the original method when employing random selection within bands. However, differences between the RTM method and the original method did not exist when employing diversity-based selection within bands. The SEM plus RTM method produced seven more majority selectees than the original method when employing the diversity-based fixed banding procedure. Furthermore, the SEM and RTM method produced eight more majority selectees than the original method when employing the fixed bands (random) procedure.

Although differences in minority selection proportions fell short of significance, the fact that employing the corrected SEM formula produces different minority selection ratios entails considerable implications. For each SED method, the diversity-based sliding bands procedure reached the target minority selection ratio. Furthermore, the diversity-based sliding bands procedure produced significantly higher minority selection ratios than standard top down, random fixed bands, and random sliding banding procedures. Overall, applying the proposed corrections produced different selection decisions than the original SED method, particularly for the diversity-based fixed banding procedure.

Buttigieg (2006) Data

Table 4 presents the descriptive statistics for scores on the Wonderlic used by Buttigieg (2006). This data set was designed to address the criticisms of Schmidt (1991), particularly that the Cascio et al. (1991) data did not represent the current state of group differences in test scores. In this sample, the standard deviation of White applicants ($s = 6.00$) is only .39 points higher than the minority standard deviation (weighted $s = 5.61$). Furthermore, these data reflect the parameters of an actual, modern sample of industrial applicants taking a commonly used selection measure.

Table 5 provides the means, lowest percentile, selection by race, minority selection ratio, and 95% confidence intervals of the minority selection ratios of the four SED methods using the Buttigieg (2006) data, selecting for 90 employees. SED methods were first compared by observing mean differences between selection decisions. Comparing the new SEM method to the original method, selection means increased for the fixed bands (random) procedure from the original ($M = 66.67$) to the new SEM ($M = 66.94$) methods. Selection means also increased for the diversity-based fixed bands procedure from the original ($M = 68.32$) to the new SEM ($M =$

69.07) methods. Means also increased for the diversity-based sliding bands procedure from the original ($M = 68.32$) to the new SEM ($M = 68.47$) methods. All procedures employing RTM produced lower selection means than procedures employing the original method.

As displayed in the "Selection by race" column, employing the new SEM method results in eight more White selectees than the original method for the diversity-based fixed banding procedure. This means that for this procedure, 23.5% of the minority applicants that were accepted using the original procedure were done so erroneously. These differences are reflected by the minority selection ratios (new SEM = 29%, original = 38%), although differences fell short of significance. Regressing each raw score to the overall mean yields different selection results than other methods. Applying only the regression to the mean yields a 47% minority selection ratio for the diversity-based fixed banding procedure, compared to 38% for the original method. The diversity-based sliding banding procedures also differ in minority selection ratio when using RTM (42%) and the original procedure (38%). Except for the diversity-based fixed banding procedure using the new SEM method, all diversity-based procedures either reached the target minority selection ratio, or the target minority selection ratio was included in the 95% confidence interval. Excluding the diversity-based fixed banding procedure using the new SEM method, all diversity-based procedures produced significantly higher minority selection ratios than random and standard top down procedures.

Table 6 presents the means, lowest percentile, selection by race, minority selection ratio, and 95% confidence intervals of the minority selection ratios of the four SED methods using the Buttigieg (2006) data, selecting for 216 employees. Comparing the new SEM method with the original method, the new SEM method produced a greater selection mean ($M = 64.62$) than the original method ($M = 63.37$) when employing the fixed bands (random) procedure. Furthermore,

the new SEM method ($M = 64.83$) produced greater means than the original method ($M = 64.69$) when employing the diversity-based fixed banding procedure. For diversity-based sliding banding procedures, the new SEM method ($M = 63.61$) also produced greater means than the original method ($M = 63.47$), suggesting that the selection decision was changed when shifting from the original method to the new SEM method. All procedures employing RTM produced lower selection means than procedures employing the original method. Concurrently, all procedures employing the new SEM and RTM produced lower selection means than the original method.

For the diversity-based fixed banding procedure, employing the new SEM method resulted in the selection of fifteen additional majority selectees instead of minority selectees, compared to the original method. These results suggest that employing the original method resulted in fifteen erroneous acceptances and rejections. Concurrently, for the diversity-based sliding banding procedure, employing the new SEM formula resulted in a difference of three additional majority selectees, and three fewer minority selectees. Regressing raw scores to the mean resulted in sixteen fewer majority selectees compared to the original method, using the diversity-based fixed banding procedure. For the diversity-based sliding banding procedure, employing RTM resulted in five fewer majority selectees compared to the original method. Accordingly, minority selection ratios differed when number of races selected differed. Excluding the diversity-based procedure employing the RTM method, each diversity-based sliding bands procedure produced significantly higher minority selection ratios than all other procedures. Concurrently, the diversity-based sliding bands procedures were the only procedures to reach the target minority selection ratio or include it in its 95% confidence interval.

CHAPTER IV: DISCUSSION

The purpose of this study was to explore the effect of employing the corrected procedure for calculating bandwidth on the selection decision, as well as the practical implications of this effect. Across both data sets, there were both similarities and differences among procedures. First of all, standard top down selection produced the highest means of applicants selected. This is to be expected, as it is consistent with the classical selection model. Furthermore, standard top down selection never produced more than a 20% minority selection ratio.

Differences among Procedures

Employing random selection within bands proved to be highly inefficient for the Cascio et al. (1991) data set. That is, random within-band selection almost always produced the lowest selection means, while always failing to meet the target minority selection ratio. Furthermore, the 95% confidence interval for the minority selection ratio never included the target minority selection ratio (47%). When employing random selection within bands using the Buttigieg (2006) data, the same effect was found for the fixed banding procedures. Both means and minority selection ratios were low. However, for the sliding banding procedure, random within-band selection produced identical or nearly identical results as standard top down selection. Selection means were higher than all other banding procedures, and minority selection ratios were either the lowest or tied for the lowest using every SED method.

Diversity-based fixed banding procedures followed similar trends across both sets of data. When selecting 90 applicants, the fixed banding procedure almost always reached the target minority selection ratio, or included it in the 95% confidence interval. However, when selecting 216 applicants, the diversity-based fixed banding procedure never reached the target minority selection ratio nor included it in the 95% confidence interval. This is likely a result of the

relationship between the number of targeted minorities and the number of selectees. That is, the selection area of the distribution contained enough minorities to satisfy the minority selection ratio for 90 selectees, but not 216 selectees. Once those minorities were selected from within the band, majority applicants had to be selected until the band was exhausted, in accordance with banding procedures.

Across all data sets, the diversity-based sliding bands procedure appeared to be the most efficient. That is, compared to other procedures, it did not appear to sacrifice considerable selection mean in order to attain the minority selection ratio. In almost all instances, the diversity-based sliding bands procedure did not produce the lowest selection mean within SED methods. Furthermore, across both sets of data and all SED methods, the target minority selection ratio was either reached or included in the 95% confidence interval.

Differences among SED Methods

Overall, employing the corrected SEM formula resulted in smaller bands than other methods. In more than one instance, utilizing the corrected formula for the standard error of measurement produced different selection results than the method proposed by Cascio et al. (1991). This is reflected in the selection means, selections by race, and minority selection ratios. As would be expected from the use of smaller bands, employing the new SEM method in the calculation of the SED resulted in higher selection means and lower minority selection ratios. However, the critical point is the fact that the new SEM method is fundamentally different from the traditional method and this engenders differences in the size of the band. This difference leads to at least one employee who was erroneously accepted, and one employee who was erroneously rejected using the incorrect procedure. With the nature of the decisions that are

being made using these selection procedures (e.g., extending/not extending employment offers),

it is critically important that the most accurate method of calculating bandwidth be employed.

Regressing each score to the mean also altered the selection decision. Because the

procedure shortened the distance between each raw score, each band was more inclusive, and

allowed for a greater minority selection ratio. This effect was displayed in the Buttigieg (2006)

data set, but not the Cascio et al. (1991) data set. Although the implications for the RTM

procedures might not be as great, the tendency of score to regress to the mean is a recognized

psychometric issue (Smith & Smith, 2005).

Overall, the results indicated that applying the new SEM formula shortened the width of

the band, and regressing the scores to the mean widened the width of the band. Only in the

Cascio et al. (1991) data set, selecting for 216 employees, did this SED method differ from the

original formula in regards to minority selection ratio. Nevertheless, appropriate procedures

should be employed when implementing a selection model.

Practical Implications

These results entail many practical implications for practitioners. According to Dudek

(1979), the traditionally accepted method for calculating SEM is mathematically flawed. The key

finding of this research is the fact that employing the corrected SEM when calculating bandwidth

produced a different selection decision than the method proposed by Cascio et al. (1991). That is,

in some cases, the original method resulted in erroneous acceptances and rejections. When SED

methods were compared using a modern industrial sample taking a popular selection device,

differences between methods became more manifest.

The first implications of this research are the potential legal ramifications. It was shown

that employing the corrected SED method produced a different selection decision than the

traditional method. If an erroneously rejected employee could prove that an organization was employing a selection method that was proven to be mathematically flawed, and that he or she was rejected due to the discrepancy between the corrected procedure and the currently employed procedure, this could be grounds for a potential lawsuit. At the very least, it would greatly undermine the credibility of the subject matter expert presenting the data and analyses.

Ethical implications must also be considered. According to the Ethical Principles of Psychologists and Code of Conduct, "Psychologists seek to promote accuracy, honesty, and truthfulness in the science, teaching, and practice of psychology" (APA, Principle C). The SEM formula proposed by Dudek (1979) should be considered because it more effectively measures an individual's true score. This is a critical objective of psychometrics, and any method proven to improve on current procedures should be considered. Employing the most accurate psychometric methods is more than just an insinuation – it a responsibility of psychologists. This is especially imperative given the gravity of employment selection decisions.

Study Limitations and Directions for Future Research

This study was not without its limitations. The first potential limitation of this research is the fact that analyses were only conducted on two sets of data. Although the addition of the data provided by Buttigieg (2006) was intended to address the criticisms of Schmidt (1991), it would likely be beneficial to replicate these procedures on other samples. This replication would add to the generalizability of the results. Another limitation is the lack of significance testing for mean differences. Although this could be an area of interest, the purpose of this research was not to explore which method was "optimal," per se, but rather to discover potential differences between SED methods. Testing for significant mean differences would suggest that one method or procedure was more effective than another. This research was strictly observational.

The lack of additional selection criteria is another limitation of this research. In practical application, when employing diversity-based procedures, selecting on race alone should not be the sole criteria of selection from within the band. This has deemed to be an illegal application of banding (Hartigan & Wigdor, 1989; Henle, 2004). In addition to race, factors such as experience and/or qualifications should be considered. However, in each data set, the only predictors that were available were the raw scores on the selection measure and race. Future research should explore potential differences between SED methods when other criteria are considered.

Finally, the inclusion of criterion data as suggested by Aguinis, Cortina, and Goldberg (1998) is an area of future research. This procedure involves calculating a band for scores on a criterion measure. According to Aguinis et al. (1998), "the ultimate goal of banding, and any other selection decision-making procedure in staffing situations, is to ascertain whether two applicants will perform at similar levels on the job" (p. 353). When available, criterion data should be included in the selection decision.

CHAPTER V: CONCLUSION

The purpose of this research was to demonstrate the effect that proper application of statistically-based banding has on selection decisions, as well as the implications this has for psychologists. As proven by Dudek (1979), the current procedure for calculating the standard error of measurement is mathematically flawed. This research demonstrates that there is a difference in the selection decisions produced by the corrected method and the traditional methods. The legal implications of this difference, the apparent controversy of statistically-based banding, and the gravity of the situation are all grounds for reexamining current statistically-based banding procedures.

REFERENCES

Aguinis, H., Cortina, H. M., & Goldberg, E. (1998). A new procedure for computing equivalence bands in personnel selection. *Human performance, 11,* 351-365.

American Psychological Association. (2003, June 1). *American Psychological Association Ethical Principles of Psychologists and Code of Conduct*. Retrieved December 2, 2009, from http://www.apa.org/ethics/code.html.

Barrett, G. V., & Lueke, S. B. (2004). Legal and practical implications of banding for personnel selection. In H. Aguinis (Ed.), *Test-score banding in human resource selection* (pp. 71-111). Westport, CT: Praeger Publishers.

Bobko, P. & Roth, P. L. (2004). Personnel selection with top-score referenced banding: On the inappropriateness of current procedures. *International Journal of Selection and Assessment, 12*, 291-298.

Bobko, P., Roth, P. L., & Nicewander, A. (2005). Banding selection scores in human resource management decisions: Current inaccuracies and the effect of conditional standard errors. *Organizational research methods, 8*, 259-273.

Bridgeports Guardians, Inc. v. City of Bridgeport, 735 F. Supp. 1126 (D. Conn. 1990), *aff'd,* 933 F.2d 1140 (2nd Cir. 1991).

Buttigieg, S. (2006). Adverse impact Study: Gender and race differences in scores on the Wonderlic Personnel Test. *Applied H.R.M. Research, 11,* 73-74.

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., & Zedeck, S. (1983). Open a new window in rational research planning: Adjust alpha to maximize statistical power. *Personnel Psychology, 36*, 517-526.

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.

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

Chicago Firefighters Local 2 v. City of Chicago, 1999 U.S. Dist. LEXIS 20310 (N.D. Ill 1999), *aff'd*, 249 F.3d (7th Cir. 2001).

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

Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory.* Mason, Ohio: Cengage Learning.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86*, 335-337.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco, CA: Freeman.

Gulliksen, H. (1950). *Theory of mental tests.* New York, NY: Wiley.

Hanges, P. J., Grojean, M. W., & Smith, D. B. (2000). Bounding the concept of test banding: Reaffirming the traditional approach. *Human Performance, 13*, 181-198.

Hartigan, J. A. & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Henle, C. A. (2004). Case review of the legal status of banding. *Human Performance, 17*, 415-432.

Krause, M. S. (2008). Regression toward the mean in effectiveness studies: Theoretically

possible, not mathematically inevitable. *Quality & Quantity: International Journal of

Methodology, 42*, 859-865.

Murphy, K. R., & Myors, B. (1995). Evaluating the logical critique of banding. *Human

Performance, 8*, 191-201.

Officers for Justice v. Civil Service Commission of San Francisco, 1992 U.S. Dist. LEXIS 3098

(N.D. Calif. 1992).

Ricci, et al. v. DeStefano ,et al., 557 U.S. 31 (2009).

Roberts, A. O. H. (1980). Regression toward the mean and the regression-effect bias. In G.

Echternacht (Guest Ed.,) *New directions for testing and measurement* (Vol. 8, pp. 59–82).

Rocconi, L. M., & Ethington, C. A. (2009). Assessing longitudinal change: Adjustment for

regression to the mean effects. *Research in Higher Education, 50*, 368-376.

Sackett, P. R. & Roth, L. (1991). A monte carlo examination of banding and rank order methods

of test score use in personnel selection. *Human Performance, 4*, 279-295.

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.

Siskin, B. R. (1995). Relation between performance and banding. *Human Performance, 8*, 215-

226.

Smith, G., & Smith, J. (2005). Regression to the mean in average test scores. *Educational

Assessment, 10*, 377-399.

Tayac v. City and County of San Francisco, 1999 U.S. App. LEXIS 31863 (9th Cir. 1999).

Zedeck, S., Outtz, J., Cascio, W. F., & Goldstein, I. L. (1991). Why do "testing experts" have

such limited vision? *Human Performance, 4*, 297-308.

APPENDIX A: TABLES

Table 1
Descriptive statistics for race group test scores (Cascio, Outtz, et al., 1991 data).

Group	N [a]	M	s
White	1803	75.6	10.4
African American	681	60.4	13.23
Hispanic	463	65.56	14.36
Asian	230	66.61	14.01
Filipino	163	60.76	14.21
Native American	35	69.14	13.43

Note. [a]Two participants did not report their race.

Table 2

Means, lowest percentile, selection by race, minority selection ratio, and confidence intervals of the five SED methods (Cascio et al., 1991 data, 90 selectees).

| Name | Mean | Lowest % | Selection by race | | | | | | Minority Selection Ratio | 95% Confidence Interval | |
			White	Black	Hispanic	Asian	Filipino	Native American		Lower Limit	Upper Limit
Original 12 point[a]											
Standard Top Down	91.62	97th	73	4	4	7	1	1	.19	.12	.29
Fixed Bands (random)	86.91	84th	74	4	5	4	2	1	.18	.11	.28
Fixed Bands (diversity-based)	90.66	87th	48	18	13	6	4	1	.47	.36	.57
Sliding Bands (random)	87.24	84th	70	3	13	3	1	0	.22	.14	.32
Sliding Bands (diversity-based)	90.66	87th	48	18	13	6	4	1	.47	.36	.57
New SEM											
Standard Top Down	91.62	97th	73	4	4	7	1	1	.19	.12	.29
Fixed Bands (random)	88.31	90th	72	5	3	8	1	1	.20	.13	.30
Fixed Bands (diversity-based)	90.92	90th	52	14	13	6	4	1	.42	.32	.53
Sliding Bands (random)	88.70	90th	73	4	8	5	0	0	.19	.12	.29
Sliding Bands (diversity-based)	90.66	87th	48	18	13	6	4	1	.47	.36	.57
Regression to Mean[b]											
Standard Top Down	89.99	97th	73	4	4	7	1	1	.19	.12	.29
Fixed Bands (random)	85.66	84th	73	4	7	3	3	0	.19	.12	.29
Fixed Bands (diversity-based)	89.61	87th	48	18	13	6	4	1	.47	.36	.57
Sliding Bands (random)	85.80	82nd	75	4	4	4	2	1	.17	.10	.26
Sliding Bands (diversity-based)	89.61	87th	48	18	13	6	4	1	.47	.36	.57
New SEM and RTM[b]											
Standard Top Down	89.99	97th	73	4	4	7	1	1	.19	.12	.29
Fixed Bands (random)	87.10	87th	73	6	7	3	0	1	.19	.12	.29
Fixed Bands (diversity-based)	89.61	87th	48	18	13	6	4	1	.47	.36	.57
Sliding Bands (random)	87.18	87th	71	5	7	7	0	0	.21	.14	.31
Sliding Bands (diversity-based)	89.61	87th	48	18	13	6	4	1	.47	.36	.57

Note. [a]Replication of Cascio et al. (1991) method. [b]Raw scores are regressed to the overall mean. Dotted line denotes target minority selection ratio (47%), based on applicant pool.

Table 3

Means, lowest percentile, selection by race, minority selection ratio, and confidence intervals of the five SED methods (Cascio et al., 1991 data, 216 selectees).

Name	Mean	Lowest %	Selection by race						Minority Selection Ratio	95% Confidence Interval	
			White	Black	Hispanic	Asian	Filipino	Native American		Lower Limit	Upper Limit
Original 12 point[a]											
Standard Top Down	89.81	92nd	172	11	12	16	3	2	.20	.15	.26
Fixed Bands (random)	87.07	84th	171	11	12	16	3	2	.21	.16	.27
Fixed Bands (diversity-based)	89.35	84th	137	22	30	15	10	2	.37	.30	.43
Sliding Bands (random)	86.55	82nd	172	10	20	8	6	0	.20	.15	.26
Sliding Bands (diversity-based)	88.61	72nd	115	44	30	15	10	2	.47	.40	.54
New SEM											
Standard Top Down	89.81	92nd	172	11	12	16	3	2	.20	.15	.26
Fixed Bands (random)	88.54	90th	176	11	16	10	1	2	.19	.14	.24
Fixed Bands (diversity-based)	89.70	90th	160	14	21	15	4	2	.26	.20	.32
Sliding Bands (random)	87.90	84th	170	10	16	13	6	1	.21	.16	.27
Sliding Bands (diversity-based)	88.61	72nd	115	44	30	15	10	2	.47	.40	.54
Regression to Mean[b]											
Standard Top Down	88.81	92nd	172	11	12	16	3	2	.20	.15	.26
Fixed Bands (random)	86.22	84th	173	10	11	13	8	1	.20	.15	.26
Fixed Bands (diversity-based)	88.37	84th	137	22	30	15	10	2	.37	.30	.43
Sliding Bands (random)	86.38	84th	170	8	18	13	6	1	.21	.16	.27
Sliding Bands (diversity-based)	87.67	72nd	115	44	30	15	10	2	.47	.40	.54
New SEM and RTM[b]											
Standard Top Down	88.81	92nd	172	11	12	16	3	2	.20	.15	.26
Fixed Bands (random)	86.89	87th	179	6	15	9	6	1	.17	.12	.23
Fixed Bands (diversity-based)	88.50	87th	144	18	30	15	7	2	.33	.27	.40
Sliding Bands (random)	86.70	87th	172	8	20	12	3	1	.20	.15	.26
Sliding Bands (diversity-based)	87.67	72nd	115	44	30	15	10	2	.47	.40	.54

Note. [a]Replication of Cascio et al. (1991) method. [b]Raw scores are regressed to the overall mean. Dotted line denotes target minority selection ratio (47%), based on applicant pool.

Table 4

Descriptive statistics for race group test scores (Buttigieg, 2006 data).

Group	N^a	N^b	M	s
White	1030	1803	21.81	6.001
African American	158	681	18.36	5.063
Hispanic	43	463	17.14	7.259
Asian	18	393	17.28	5.529
Native American	28	35	18.39	6.184

Note. [a]Original sample size from Buttigieg (2006) data. bAnalyzed sample size used based on Cascio, Outtz et al. (1991) data.

Table 5

Means, lowest percentile, selection by race, minority selection ratio, and confidence intervals of the five SED methods (Buttigieg, 2006 data, 90 selectees).

			Seclection by race					Minority Selection Ratio	95% Confidence Interval	
Name	Mean	Lowest %	White	Black	Hispanic	Asian	Native American		Lower Limit	Upper Limit
Original 12 point[a]										
Standard Top Down	69.74	97th	79	0	9	1	1	.12	.07	.21
Fixed Bands (random)	66.67	91st	77	4	4	3	2	.14	.08	.24
Fixed Bands (diversity-based)	68.32	91st	56	13	13	7	1	.38	.28	.49
Sliding Bands (random)	69.74	97th	79	0	8	1	2	.12	.07	.21
Sliding Bands (diversity-based)	68.32	91st	56	13	13	7	1	.38	.28	.49
New SEM										
Standard Top Down	69.74	97th	79	0	9	1	1	.12	.07	.21
Fixed Bands (random)	66.94	93rd	77	2	9	1	1	.14	.08	.24
Fixed Bands (diversity-based)	69.07	93rd	64	8	13	4	1	.29	.20	.40
Sliding Bands (random)	69.74	97th	80	0	8	1	1	.11	.06	.20
Sliding Bands (diversity-based)	68.47	91st	56	13	13	7	1	.38	.28	.49
Regression to Mean[b]										
Standard Top Down	66.17	97th	79	0	9	1	1	.12	.07	.21
Fixed Bands (random)	63.52	87th	76	1	10	2	1	.16	.10	.24
Fixed Bands (diversity-based)	63.99	87th	48	18	13	10	1	.47	.36	.57
Sliding Bands (random)	66.17	97th	79	0	8	1	2	.12	.07	.21
Sliding Bands (diversity-based)	64.49	89th	52	17	13	7	1	.42	.32	.53
New SEM and RTM[b]										
Standard Top Down	66.17	97th	79	0	9	1	1	.12	.07	.21
Fixed Bands (random)	63.69	91st	81	1	6	2	0	.10	.05	.19
Fixed Bands (diversity-based)	64.92	91st	56	13	13	7	1	.38	.28	.49
Sliding Bands (random)	66.17	97th	79	0	8	1	2	.12	.07	.21
Sliding Bands (diversity-based)	64.92	91st	56	13	13	7	1	.38	.28	.49

Note. [a]Replication of Cascio et al. (1991) method. [b]Raw scores are regressed to the overall mean. Dotted line denotes target minority selection ratio (47%), based on applicant pool.
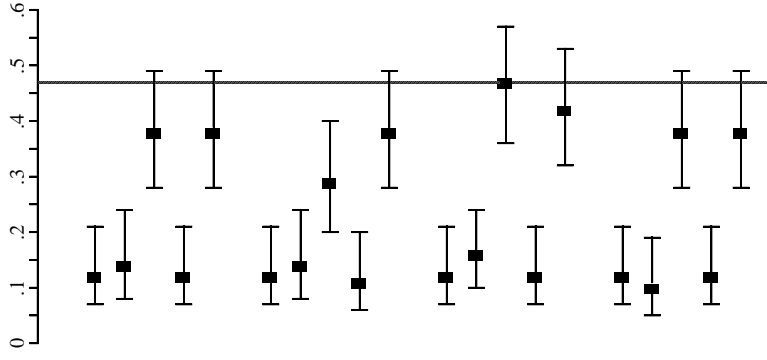
Table 6

Means, lowest percentile, selection by race, minority selection ratio, and confidence intervals of the five SED methods (Buttigieg, 2006 data, 216 selectees).

| Name | Mean | Lowest % | Selection by race | | | | | Minority Selection Ratio | 95% Confidence Interval | |
			White	Black	Hispanic	Asian	Native American		Lower Limit	Upper Limit
Original 12 point[a]										
Standard Top Down	64.83	93rd	175	8	22	4	7	.19	.14	.25
Fixed Bands (random)	63.37	91st	173	9	20	4	10	.20	.15	.26
Fixed Bands (diversity-based)	64.69	91st	160	13	30	7	6	.26	.20	.32
Sliding Bands (random)	64.83	93rd	179	8	19	3	7	.17	.12	.23
Sliding Bands (diversity-based)	63.47	82nd	121	42	30	17	6	.44	.37	.51
New SEM										
Standard Top Down	64.83	93rd	175	8	22	4	7	.19	.14	.25
Fixed Bands (random)	64.62	93rd	176	8	22	4	6	.19	.14	.24
Fixed Bands (diversity-based)	64.83	93rd	175	8	22	4	7	.19	.14	.25
Sliding Bands (random)	64.81	92nd	178	8	20	3	7	.18	.13	.23
Sliding Bands (diversity-based)	63.61	82nd	124	39	30	17	6	.43	.36	.49
Regression to Mean[b]										
Standard Top Down	61.85	93rd	175	8	22	4	7	.19	.14	.25
Fixed Bands (random)	59.84	87th	171	13	24	5	3	.21	.16	.27
Fixed Bands (diversity-based)	61.44	87th	144	25	30	10	7	.33	.27	.40
Sliding Bands (random)	61.85	93rd	177	8	21	3	7	.18	.13	.24
Sliding Bands (diversity-based)	60.42	80th	116	44	30	20	6	.46	.40	.53
New SEM and RTM[b]										
Standard Top Down	61.85	93rd	175	8	22	4	7	.19	.14	.25
Fixed Bands (random)	60.49	91st	173	8	21	6	8	.20	.15	.26
Fixed Bands (diversity-based)	61.72	91st	159	13	30	7	7	.26	.21	.33
Sliding Bands (random)	61.85	93rd	180	8	18	3	7	.17	.12	.22
Sliding Bands (diversity-based)	60.65	82nd	121	42	30	17	6	.44	.37	.51

Note. [a]Replication of Cascio et al. (1991) method. [b]Raw scores are regressed to the overall mean. Dotted line denotes target minority selection ratio (47%), based on applicant pool.