Spring 2022

# A Monte Carlo Analysis of Standard Error-Based Methods in the Construction of Standard Error of Difference Bandwidths

Diljot Singh Kochhar

# A MONTE CARLO ANALYSIS OF STANDARD ERROR-BASED METHODS IN THE CONSTRUCTION OF STANDARD EORROR OF DIFFERNECE BANDWIDTHS

A Thesis Submitted in Partial Fulfillment
Of the Requirements for the Degree
Master of Science

Department of Psychological Sciences
Western Kentucky University
Bowling Green, Kentucky

By
Diljot Singh Kochhar

May 2022

A MONTE CARLO ANALYSIS OF STANDARD ERROR-BASED METHODS IN THE CONSTRUCTION OF STANDARD EORROR OF DIFFERNECE BANDWIDTHS

April 12, 2022
Defense Date

Dr. Reagan Brown
Committee Chair

Dr. Katrina Burch
Committee Member

Dr. Aaron Wichman
Committee Member

Ranjit T. Koodali

Associate Provost for Research & Graduate Education

ABSTRACT

A MONTE CARLO ANALYSIS OF STANDARD ERROR-BASED METHODS IN THE CONSTRUCTION OF STANDARD EORROR OF DIFFERNECE BANDWIDTHS

The purpose of this study is to examine the if the standard error of estimate (SEE) is more effective than the standard error of measurement (SEM) when used in the construction of use the standard error of the difference (SED)-based bandwidths. It was hypothesized that the standard error of estimate would be a more effective equation to use than the standard error or measurement because it allows for the estimation of a range of true scores around an observed score instead of the opposite scenario. To examine the effectiveness of the equations, a Monte Carlo analysis was employed to determine the effectiveness of the bandwidths at capturing score differences. Results indicate that the SEE-based bandwidth led to an accurate as well as a more efficient bandwidth than the SEM-based bandwidth.

I would like to dedicate this to my parents. Thank you for always supporting me through all my endeavors in life. Also, a special thank you to Markel Litvinchuk one of the greatest men I've had the pleasure to know.

ACKNOWLEDGMENTS

I would like to extend a special thank you to my thesis chair and professor Dr. Reagan Brown. Thank you for being one of the best professors I've ever had the pleasure to have and for helping me fall in love with statistics. Your passion for teaching and pam risk related jokes will always be something I cherish. I would also like to thank Dr. Katrina Burch and Dr. Aaron Wichman for being on my committee and making my time as a graduate student something that I will always look back on in fondness. Thank you!

Table Of Contents

LIST OF TABLES

## Introduction

The effective use of tests scores to make informed employment decisions for selection and promotion is central to the field of industrial-organizational psychology. This activity is not only important in the sense that it is good practice but also has a major impact on organizational functioning and can lead to serious issues (e.g., legal and regulatory violations) if done improperly (Gasperson, Bowler, Wuensch, & Bowler, 2013). Proper selection decisions also allow organizations to ensure that they are selecting candidates that will truly perform better than other candidates once hired and not on the basis of possible errors that may have occurred during the testing process.

### Classical Test Theory

Classical Test Theory (CTT) provided the original basis for psychologists to assess measurement precision, principally through the concept of reliability (Lord & Novick, 1968). CTT allows psychologists to investigate the relationship between scores that are seen on tests and what is not seen (Spearman, 1904). Specifically, CTT defines an observed score ($X$) as the sum of a stable component, known as the true score ($T$), and an unstable component, known as error ($E$).

**True Score:** The true score in CTT is the stable component of the test score; it has no random error in it ($T = X - E$). In CTT, the true score is assumed to be constant for a test taker across multiple test administrations. In contrast, the observed score varies across administrations due to random error. $T$ is the expected value of $X$ over multiple administrations of a test to a test taker (i.e., $E(X) = T$; Equation 2.3.1, Lord & Novick, 1968, p. 30) because $X = T + E$, and the

expected for E is zero (thus, $E(X) = T + 0$). Finally, if there is no random error in a measure (i.e., $E = 0$ for all test takers) $X = T$; in other words, the observed score is the true score.

**Error Score:** $E$, which is random error (Lord & Novick, 1968) is a variable that which, by definition, has no stable or systematic part to it. The nature of $E$ is established by the following assumptions. Among the individuals taking the test, $E$ is uncorrelated with $T$ ($r_{et} = 0$, Equation 2.7.1b, Lord & Novick, 1968, p. 36). Second, scores on $E$ on a parallel test will not be correlated among test takers ($r_{e1e2} = 0$; Equation 2.7.1d, Lord & Novick, 1968, p. 36). Last, the expected scores for E on the same test administered multiple times to the same test taker is zero ($E(E) = 0$; Equation 2.4.2, Lord & Novick, 1968, p. 31). Depending on the cause of the random error, a test taker's score could either be affected in a way that positively impacts observed scores or negatively impacts observed scores. For example, if an individual does not know the answer to a question and randomly guesses the correct answer, this response results in an error that positively impacted the test taker's observed score such that the observed score is higher than the true score. However, if the test taker accidently makes an incorrect answer, this error would negatively impact the test taker's observed score, making it lower than the true score.

**Systematic error in classical test theory:** Systematic error is error that occurs in the same way over time in a repeated measure (Guion, 1965). A couple examples of systematic errors on a measure include rating biases (for rating data) and test wiseness (for optimal performance items).

It is a fundamental limitation of CTT that systematic error is not addressed. Moreover, a test may be without random error (i.e., $E = 0$ for all test takers) but still contain systematic errorsof various kinds. These systematic errors are part of the true scores in the CTT model. Guion

(1965) reconceptualized the original $X = T + E$ equation as: $X = s + e$ where s is a systematic component, a composite of true measure and any constant (i.e., nonrandom) error.

Although it would be desirable to have random error and systematic error in the equation, leaving $T$ to be completely without error, the assumptions of CTT do not allow for this. It is important to note that it is not the label of the term that defines the characteristics of those terms; rather the assumptions of those terms that define the characteristics of the terms.

**Reliability:** Reliability describes the amount of random error in a measure. CTT defines reliability as the ratio of true score variance (i.e., variance that is not random) to observed score variance (i.e., total variance).

$$S^2_X = S^2_T + S^2_E \quad (1)$$

The reliability coefficient gives information on how much of the observed score variance is due to the systematic factors and not the random error present in the measure. It is important to note that the reliability coefficient only allows us to view the reliability of a measure as a whole, not individual scores (Harvill, 1991). Reliability can be estimated through a variety of methods such as the test-retest and internal consistency reliability. Once an estimate of the reliability coefficient has been obtained for a measure, one can take further steps to determine likely magnitude of random error in individual scores.

**Confidence Intervals and True Scores**

Because true scores are unknowable in practice, we can only estimate the likely range of values for a true score given an observed score. A confidence interval is used for this estimation. The reportage of the likely location of a test taker's true score allows unreliability to be expressed in nontechnical terms for the uninitiated consumer of statistical information (Harvill, 1991). Confidence intervals are formed usually at a 95% level, but any desired value is available.

These confidence intervals are computed with a standard error statistic. This standard error is based on the reliability and standard deviation of the test but can take several forms. Because there are multiple relevant standard error equations, differing slightly in their construction, the resultant confidence intervals have different ranges.

**Standard Error of Measurement**

The standard error of measurement (SEM) is the standard deviation of observed scores for a given true score and is determined by the reliability and standard deviation of the test (Dudek, 1979). Stated differently, the SEM is the standard error of observed scores that would be observed if the test taker were to take the test an infinite number of times (Cascio et al., 1991).

$$SEM = S_X\sqrt{1 - r_{XX}} \quad (2)$$

Thus, reliability is directly related to the standard error of measurement. Greater levels of reliability result in lower values for the SEM. The difference between the SEM and the reliability coefficient is that the SEM allows us to view the effects of random error at individual score levels whereas the reliability coefficient only allows us to view random error at the test level (Ghiselli, Campbell, & Zedeck, 1981).

An SEM-based confidence interval appears to allow for a useful applied use of information regarding random error (i.e., reliability) at the individual score level. The SEM is often misused when forming these confidence intervals. This misuse results from the design of the SEM equation; the SEM equation for confidence intervals forms observed scores around a constantly held true score (Dudek, 1979). However, the common purpose of the confidence intervals in practice is to start with a known observed score and form an interval around it to determine the likely location of the unknown true score. Thus, the SEM is effectively useless in

applied practice and confidence intervals resulting from this misuse should not be interpreted as indicating the likely location of the test taker's true score.

**Standard Error of Estimate**

The standard error of estimate (SEE) allows for the creation of an index of error around an observed score, giving the likely location of the true score (Dudek, 1979; Lord & Novick, 1968). The resultant confidence interval matches what is needed in applied practice: the known observed score is used as the starting point to determine the likely location of the unknown true score. The equation for the SEE is similar to that of the SEM but with a slight adjustment.

$$SEE = S_X\sqrt{r_{XX}(1 - r_{XX})} \quad (3)$$

The difference between the SEE and the SEM equations is the multiplication of the reliability coefficient after subtracting the reliability equation from one.

Dudek (1979) stated that an adjustment to the mean for the observed score should be used when creating the confidence intervals. This adjustment moves the observed score closer to the mean and this effectively makes score less extreme. The logic for this adjustment is that because $X = T + E$, where $E$ is a random variable, extreme values for $X$ are likely in part the result of extreme values for $E$. Adjusting the observed score closer to the mean offers a better starting point for the confidence interval. The regression to the mean is calculated by subtracting the observed score form the mean, multiplying it by the reliability coefficient, and adding back the mean to this adjusted value.

$$X_{RTM} = \bar{X} + r_{XX}(X - \bar{X}) \quad (4)$$

**Research on SEM and SEE-Based Confidence Intervals**

Wichert (2020) used a Monte Carlo analysis to test four methods for constructing a confidence interval around an observed score with the goal of accurately locating the true score.

Four types of confidence intervals were investigated: SEM and SEE-based confidence intervals with and without a regression to the mean adjustment (RTM) to the observed score. Her results indicated that both the SEM (without a regression to the mean adjustment) and the SEE (with a regression to the mean adjustment) based intervals were able to correctly locate the true score 95% of the time with the 95% confidence interval. However, although the SEM does create a 95% confidence interval successfully, it was not as narrow as the SEE equation. Other factors equal, a narrower interval is to be preferred as it reduces ambiguity regarding the location of the true score. Thus, Wichert's (2020) results support Dudek's (1979) recommendation regarding confidence intervals around an observed score.

**Standard Error of the Difference**

The Standard Error of Difference (SED) allows researchers to compare two observed scores to determine whether the true scores are "reliably different" (Casico et al., 1991, p. 241). That is, for tests with reliabilities less than 1.0 (i.e., all tests in existence) two test takers with the same true score will likely have different observed scores simply due to random error. The SED offers an index of the expected difference between observed scores due to random error when the true scores are the same. If a researcher chooses a 95% confidence interval, then (the absolute value of) any difference between observed scores greater than 1.96 × SED exceeds that which would be observed between two identical true scores 95% of the time (i.e., in Cascio et al., 1991, parlance, the test scores are "reliably different"). In summary, the SED can be used to compare two observed scores to determine whether the observed difference is likely due to random error.

Comparisons of observed score differences to the SED are not mere academic exercises. In personnel selection the SED has a natural application in test score banding (Cascio et al., 1991). Banding is an alternative to ranking methods that are often used during selection

procedures (Cascio et al., 1991). “The fundamental purpose of statistically based banding is to generate a range of possible true scores around an observed score so that scores within that band can be equated with one another” (Gasperson et al., 2013, p.46). The proper use of equations for confidence intervals involving CTT is crucial for practical applications and can have real impact on decisions (e.g., selection) made in organizational settings.

The SED, as defined by Gulliksen (1950), is calculated by multiplying the SEM by the square root of two. However, the use of the SEM in the SED equation may be inappropriate for the purpose for which the SED is used. As with the case with confidence intervals around an observed score, the use of the SEM in this equation may be misplaced as the SEM is designed for intervals around a true score (Dudek, 1979). To the contrary, the SEE, which is designed for intervals around an observed score, should be the standard error statistic of choice. A second consideration is that the SEE creates a confidence interval of true scores around an observed score more efficiently than the SEM (a narrower interval).

**The Present Study**

The present study uses a Monte Carlo analysis to investigate the effectiveness of the SEM and SEE (with and without a regression to the mean adjustment) in SED-based band construction. The effectiveness of these procedures will be investigated for two-tailed as well as one-tailed banding procedures where the null is true. Finally, I will also investigate the effectiveness of one-tailed bands where the null is false.

A Monte Carlo analysis is a technique that uses large data sets to test hypotheses regarding the effectiveness of analytical procedures in a variety of conditions. This technique allows for the researcher to manipulate large data sets in a variety of ways (e.g., adjusting reliability levels) to determine the conditions under which various statistical procedures are most

effective. For the purposes of this study, a Monte Carlo analysis will allow for the examination of the accuracy of different methods for producing a confidence interval for true scores around an observed score using various forms of the SED equation.

Based on previous research of SEM and SEE-based confidence intervals by Wichert (2020), I expect to see that SED employing the SEE equation with the observed scores regressed to the mean will produce optimal results. The SEE bands will achieve desired accuracy (e.g., the true scores will be in the band 95% of the time for the 95% confidence interval) while producing a band that is narrower than ones achieved from SEM-based SED.

## Method

### Design

A Monte Carlo analysis was used to determine the accuracy of the SEE and SEM-based SED equation with and without a regression to the mean adjustment to the observed scores. Thus, there were four SED-based bands developed for each condition in this study: SEM-based SED band with and without the regression to the mean and an SEE-based SED band with and without the regression to the mean for both the one-tailed and two-tailed banding procedures.

For the two-tailed banding procedure, I compared observed scores for which the true scores were the same. Accuracy of the SED equations was assessed by whether the comparison of the observed scores to the SED (i.e., whether the observed score difference is within the computed SED) matched the reality of no difference between the true scores. Because I computed the SED for a 95% interval, the comparisons should yield the conclusion of no difference in 95% of the comparisons.

For the one-tailed banding procedure, I looked at two different conditions. In the condition where the null is true, the true scores were identical. As with the two-tailed test, if the

difference between the observed scores was less than the SED, then the scores would be correctly classified as not different. As before, this conclusion should be reached 95% of the time for a 95% interval. For the condition where the null hypothesis was false, the true scores were different. When the observed difference was greater than the SED, then we conclude that the true score is greater.

Aside from the four methods of SED computation, the only other manipulated variable was the reliability of the measure. Reliability was ranged from .1 to .9, increasing in .1 increments. True scores were generated from a distribution of standardized data having a mean of zero and a standard deviation of 1.0. The standard deviation of the observed scores varied as a function of the reliability. For each of the nine reliability conditions, the procedure was repeated 1,000,000 times and results were averaged across these 1,000,000 replications.

## Results

Table 1 displays the percent of observations for which the observed score difference fell within the bandwidth for the two-tailed, null true condition. Because the null is true (i.e., identical true scores) for this condition, the observed score difference should fall within a 95% band 95% of the time. The SEM and SEE-RTM-based SED bands produced the most accurate results with the bands accurately capturing the observed score differences within the band about 95% of the time on average across all reliabilities. The least accurate band was the SEE-based band which only accurately captured the score differences within the band 78% on average across reliabilities. The SEM-RTM accurately captured score differences within the band about 88% of the time on average across the reliabilities. The results for the SEM-based and SEE-RTM-based SED bands are identical because the observed score mean was set to zero. This issue will be addressed further in the discussion.

**Table 1**

*SEM-based and SEE-based SED Band Performance for Two-Tailed Null True Condition*

| Reliability | SEM | SEE | SEM-RTM | SEE-RTM |
|---|---|---|---|---|
| .1 | 95.018 | 46.461 | 100.000 | 95.018 |
| .2 | 94.984 | 61.936 | 99.999 | 94.984 |
| .3 | 94.994 | 71.635 | 99.963 | 94.994 |
| .4 | 95.013 | 78.514 | 99.798 | 95.013 |
| .5 | 95.020 | 83.441 | 99.449 | 95.020 |
| .6 | 94.972 | 87.049 | 98.853 | 94.972 |
| .7 | 94.999 | 89.900 | 98.092 | 94.999 |
| .8 | 94.936 | 91.983 | 97.146 | 94.936 |
| .9 | 94.987 | 93.695 | 96.103 | 94.987 |

*Note.* Table entries are the mean values across one million replications. Table entries indicate percent of observations for which the observed score difference is less than bandwidth.

Table 2 displays results for the one-tailed, null true condition. Because directional comparisons of observed scores are formed post hoc (the designation of bigger and smaller observed score are not set a priori), the one-tailed band must be formed using a *z* value of 1.96, making these results identical to a two-tailed band. As with Table 1 the percentage of time the band accurately captured the score is shown for all four methods purposed for SED construction. Both the SEM and SEE-RTM bands were the most accurate as they captured the observed scores differences in the band about 95% of the time. The least accurate band was the SEE-based band which only accurately captured the observed score differences within the band about 78% of the time. the SEM-RTM band accurately captured the observed score differences about 88% of the time.

**Table 2**

*SEM-based and SEE-based SED Band Performance for One-Tailed Null True Condition*

| Reliability | SEM | SEE | SEM-RTM | SEE-RTM |
|---|---|---|---|---|
| .1 | 94.996 | 46.523 | 100.000 | 94.996 |
| .2 | 94.995 | 61.900 | 99.999 | 94.995 |
| .3 | 94.991 | 71.669 | 99.965 | 94.991 |
| .4 | 94.990 | 78.440 | 99.808 | 94.990 |
| .5 | 95.017 | 83.437 | 99.448 | 95.018 |
| .6 | 95.029 | 87.073 | 98.872 | 95.029 |
| .7 | 94.980 | 89.883 | 98.085 | 94.980 |
| .8 | 94.991 | 92.030 | 97.154 | 94.991 |
| .9 | 94.998 | 93.703 | 96.123 | 94.998 |

*Note.* Table entries are the mean values one million replications. Table entries indicate percent of observations for which the observed score difference is less than bandwidth.

The results for the one-tailed, null false condition are shown in Table 3. Because the null is false (and these are continuous variables), the true scores are never equal. As a result, we would not expect the observed score difference to fall within the band 95% of the time. How often the observed score difference should exceed the bandwidth for a one-tailed test is unknown and is explored in this analysis. (Ideally, a two-tailed band should never contain the observed score difference when the null is false. Any instance in which the bandwidth is greater than the observed score difference is analogous to a Type II error.) Table 3 shows results of how often the difference of the observed scores is captured in the bands for the nine levels of reliability. On average, the SEM-based and SEE-RTM-based bandwidths had the highest percentage (78%) whereas the SEE-based bandwidth had the lowest rate. This rate in which the observed score difference exceeded the bandwidth varied by reliability for all four versions of the bandwidth equation; the bands contained fewer of the observed score differences as reliability increased.

**Table 3**

*SEM-based and SEE-based SED Band Performance for One-Tailed Null False Condition*

| Reliability | SEM | SEE | SEM-RTM | SEE-RTM |
|---|---|---|---|---|
| .1 | 93.662 | 44.369 | 100.000 | 93.662 |
| .2 | 92.041 | 56.687 | 99.992 | 92.041 |
| .3 | 89.842 | 62.978 | 99.715 | 89.842 |
| .4 | 87.123 | 66.345 | 98.345 | 87.123 |
| .5 | 83.429 | 67.374 | 94.954 | 83.429 |
| .6 | 78.476 | 66.271 | 89.047 | 78.476 |
| .7 | 71.714 | 63.109 | 80.079 | 71.714 |
| .8 | 61.934 | 56.692 | 67.298 | 61.934 |
| .9 | 46.484 | 44.381 | 48.659 | 46.484 |

*Note.* Table entries are the mean values across one million replications. Table entries indicate percent of observations for which the observed score difference is less than bandwidth.

Table 4 shows the results from an extended investigation of the one-tailed, null false condition. For this analysis, I examined only the cases in which the observed score difference exceeded the bandwidth, a result that leads to the conclusion that the person with the greater observed score also has the greater true score. For these cases, I computed the percent of observations where the true score was in fact greater (in other words, the band worked as it should). As Table 4 shows, the SEE-based and SEM-RTM-based bands were the most accurate at correctly identifying cases that were out of the bandwidth and the larger observed score with an accuracy of about 95% of the time. The least accurate band was the SEE-based band which had an accuracy of about 88%. Furthermore, as long as reliability was at least .5, then all four band equations rendered the correct conclusion that the bigger observed score had the bigger true score at least 90% of the time (95% for all but the CI see-based band). Also shown in Table 4 is the mean absolute true score differences in these conditions; differences were approximately one standard deviation.

**Table 4**

*One-Tailed Null False Percent of Cases Out of Band Correctly Classified*

| Reliability | $\lvert True_1 - True_2 \rvert$ | SEM | SEE | SEM-RTM | SEE-RTM |
|---|---|---|---|---|---|
| .1 | 1.12916 | 77.22 | 65.49 | 60.35 | 77.22 |
| .2 | 1.1278162 | 85.63 | 74.37 | 100.00 | 85.63 |
| .3 | 1.1303315 | 90.53 | 81.81 | 98.14 | 90.53 |
| .4 | 1.1281557 | 93.82 | 87.45 | 98.46 | 93.82 |
| .5 | 1.1277510 | 95.93 | 91.95 | 98.79 | 95.93 |
| .6 | 1.1278838 | 97.28 | 94.97 | 98.96 | 97.28 |
| .7 | 1.1278261 | 98.37 | 97.17 | 99.20 | 98.37 |
| .8 | 1.1286949 | 99.02 | 98.55 | 99.40 | 99.02 |
| .9 | 1.1277339 | 99.53 | 99.42 | 99.64 | 99.53 |

*Note.* $\lvert True_1 - True_2 \rvert$ is the absolute value of the true score difference. Table entries are restricted to cases for which the observed score difference exceeded the bandwidth and indicate the percent of those cases where the greater observed score had the greater true score.

Table 5 shows the bandwidth for SEM-based and SEE-based SED bandwidths (95% confidence) at the nine levels of reliability. As the reliability increased, the size of the bandwidth decreased. In addition, at higher levels of reliability the magnitude of the difference between bandwidths also decreased (as is indicated by the equations for each). This table also shows the observed standard deviation of the scores captured for the SEM- and SEE-based data as well.

**Table 5**

*Bandwidth*

| Reliability | Observed SD | SEM | SEE |
|---|---|---|---|
| .1 | 3.162 | 8.315 | 2.630 |
| .2 | 2.236 | 5.544 | 2.479 |
| .3 | 1.826 | 4.234 | 2.319 |
| .4 | 1.581 | 3.395 | 2.147 |
| .5 | 1.414 | 2.772 | 1.960 |
| .6 | 1.291 | 2.263 | 1.753 |
| .7 | 1.195 | 1.815 | 1.518 |
| .8 | 1.118 | 1.386 | 1.240 |
| .9 | 1.054 | 0.924 | .877 |

*Note:* 95% Bandwidth (i.e., z = 196). True score standard deviation = 1.0; observed score standard deviation varies as a function of reliability.

## Discussion

The SED allows for the comparison of two scores to determine if two observed scores are "reliably different" (Cascio et al., 1991, p. 241). As stated by Dudek (1979), the standard error of estimate (SEE) allows for the creation of an index of error around an observed score, giving the likely location of the true score (instead of an observed score around a true score as per the SEM). The purpose of this study was to determine if an SEE-based SED bandwidth would be more effective than the traditional SEM-based SED bandwidth. In addition, I also explored the role of the regression to the mean adjustment in the efficacy of various SED bandwidths.

The results that have been produced in this Monte Carlo study show that the SEM-based and SEE-RTM based bandwidths were effective at capturing differences between scores when true scores were not different approximately 95% of the time whereas the SEE-based bandwidth was only able to accurately capture scores about 88% of the time. However, when examining the bandwidths (Table 5), it is clear to see that the bandwidth is larger for the SEM-based bandwidth then it is for the SEE-based bandwidth. The goal of the bandwidth should be accuracy as well as having the narrowest bandwidth as possible and the SEE provides that.

It is important to note that the SEM and SEE-RTM bands produced identical scores in each condition (e.g., null true, null false, one-tailed, two-tailed) at every level of reliability. Upon examination, it was discovered that a specific characteristic (observed score means were set to zero) of the study rendered algebraically equivalent equations for SEM-based and SEE-RTM-based bands (the extra reliability in the SEE-based SED was offset by the extra reliability in the regression to the mean adjustment). These functionally equivalent equations led to identical performance for the two banding methods. To further explore this issue, a follow-up analysis was conducted in which the observed score means were allowed to be non-zero values; bandwidth

performance for the SEM-based and SEE-RTM-based bands was comparable (approximately 95%) but not identical.

Another finding from the study was that even when a researcher is thinking of band comparisons in a directional (i.e., one-tailed) sense, there is not a true one-tailed test. Because the choice of observed scores is not made a priori, any comparison of the bigger observed score to the smaller observed score is effectively a two-tailed comparison and requires a z value of 1.96 for the 95% SED bandwidth to correctly classify matters (null true) 95% of the time. This is not to say that there isn't value to directional thinking (e.g., Does the person with the greater observed score actually have the greater true score?). It merely calls for an adjustment to the banding equation.

**Similarities in the SEE-RTM and SEM Monte Carlo Equation**

Upon inspection of the results, I discovered that the percentages listed in Tables 1-4 were identical for the SEM bandwidths and SEE-RTM bandwidths. Upon further analysis, I discovered that a characteristic of the Monte Carlo analysis rendered the equations algebraically identical.

In short, true and observed scores were generated to have a mean of zero. This mean of zero allowed a reliability to be factored out of the RTM equation, which then cancelled out a reliability in the SEE equation, rendering it identical to the SEM equation without a RTM adjustment, which rendered the results identical. This process is described below.

To begin, the SEE equation is compared to the RTM adjusted difference between the observed scores.

$$z \times S_X\sqrt{r_{XX}(1 - r_{XX})} > \bar{X} + \sqrt{r_{XX}}((\overline{X_1} - X_2) - \bar{X})$$

Upon factoring out the reliability from the SEE equation and substituting zero for the observed score mean in the RTM adjustment (and simplifying), we have the following.

$$z \times S_X \sqrt{(1 - r_{XX})} > ((X_1 - X_2))$$

This equation equals the comparison with the SEM-no RTM and is the reason the results that were produces were identical for the SEM and SEE-RTM bandwidth.

Once the cause of this problem was identified, I executed some analyses where the mean observed score was allowed to be non-zero values, and the results for the SEM and SEE-RTM bandwidths were no longer identical. They did produce similar results centering on approximately 95% correct classifications.

**Practical Implications**

Personnel selection is not only important for organizations but is also an essential function in the everyday role of I-O Psychologists so it is crucial that the selection procedure is undertaken in the most effective way possible. Diversity and inclusion has become a major emphasis for organizations. The benefits of diversity and inclusion far exceed mere public relations; it allows organizations to bring in different perspectives which in turn allow for the development of results for a variety of perspectives. The use of banding procedures allows for hiring managers to determine which individuals have abilities that are very similar. This determination then allows organizations to look at other factors such as age, gender, and race as selection criteria which will help with diversity and inclusion initiatives. It will also allow the avoidance of adverse impact charges which can have major legal implications for organizations.

## Conclusions

As indicated by the results, it is important to have a bandwidth that accurately identifies when differences between scores are within a range expected given the reliability of the test

while being narrow enough to not classify truly different scores as similar; the SEE-based bandwidth provides that. The benefits go beyond surface level implications and has the potential to positively impact many lives in organizations across the world.

## References


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.

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

Gasperson, S. M., Bowler, M. C., Wuensch, K. L., & Bowler, J. L. (2013). A statistical correction to 20 years of banding. *International Journal of Selection and Assessment, 21*, 46-56.

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

Guion, R. M. (1965). *Personnel testing.* New York, NY: McGraw-Hill.

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

Harvill, L. M. (1991). An NCME instructional module on standard error of measurement. *Educational Measurement: Issues and Practice, 10*, 33-41.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Wichert, E. (2020). *A Monte Carlo analysis of standard error-based methods for computing confidence intervals* (Publication No. 3203) [Master's thesis, Western Kentucky University]. https://digitalcommons.wku.edu/theses/3203