Spring 2020

A Monte Carlo Analysis of Standard Error-Based Methods for Computing Confidence Intervals

Elayna Wichert
Western Kentucky University, elayna.wichert901@topper.wku.edu

A MONTE CARLO ANALYSIS OF STANDARD ERROR-BASED METHODS FOR COMPUTING CONFIDENCE INTERVALS

A Thesis
Presented to
The Industrial-Organizational Faculty in the Department of Psychological Sciences
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

By
Elayna Wichert

May 2020

A MONTE CARLO ANALYSIS OF STANDARD ERROR-BASED METHODS FOR COMPUTING CONFIDENCE INTERVALS

Date Recommended May 20, 2020

Reagan Brown Digitally signed by Reagan Brown Date: 2020.05.26 13:44:43 -05'00'

Dr. Reagan Brown, Director of Thesis

Elizabeth Shoenfelt Digitally signed by Elizabeth Shoenfelt Date: 2020.05.26 13:49:35 -05'00'

Dr. Elizabeth L. Shoenfelt

Katrina Burch Digitally signed by Katrina Burch Date: 2020.05.26 20:18:11 -05'00'

Dr. Katrina A. Burch

Cheryl D Davis Digitally signed by Cheryl D Davis Date: 2020.06.10 16:14:00 -05'00'

Dean, Graduate School Date

I dedicate this thesis to Dr. Kevin Brown, who sparked my initial interest in this field and whose continuous support guided me to where I am today.

TABLE OF CONTENTS

LIST OF TABLES

A MONTE CARLO ANALYSIS OF STANDARD ERROR-BASED METHODS FOR COMPUTING CONFIDENCE INTERVALS

Elayna Wichert May 2020 19 Pages

Directed by: Reagan D. Brown, Katrina A. Burch, and Elizabeth L. Shoenfelt

Department of Psychological Sciences Western Kentucky University

The objective of this study is to empirically test existing techniques to calculate the likely range of values for a Classical Test Theory true score given an observed score. The traditional method for forming these confidence intervals has used the standard error of measurement (SEM) as the basis for this confidence interval. An alternate equation, the standard error of estimate (SEE), has been recommended in place of the SEM for this purpose, yet it remains overlooked in the field of psychometrics. It is important that the correct equation be used in various applications in personnel psychology. Monte Carlo analyses were conducted to investigate the performance of the various methods for computing a confidence interval around an observed score. Results indicated that the SEE equation used with an observed score regressed to the mean most accurately and efficiently located an individual's true score.

Introduction

The interpretation of test scores is central to many functions within the field of industrial-organizational (I-O) psychology. Such decisions may concern an organization's selection, promotion, training, development, and performance management procedures. Given the nature of the decisions being rendered, it is crucial that researchers and practitioners apply accurate methods to maximize the benefits of this research and practice (Gasperson, Bowler, Wuensch, & Bowler, 2013).

Classical Test Theory

Classical Test Theory (CTT), the oldest theory of measurement (Spearman, 1904), provides a framework for the interpretation and development of psychological test and assessment reliability (Lord & Novick, 1968). CTT explores the relationship between information that can be gathered from observation and information that is unobservable (Spearman, 1904). Most notably, the theory states that an individual's observed score (X) equals the sum of the hypothetical true score (T) plus measurement error (E).

$$X = T + E$$

Observed score. The observed score, X, is a random variable consisting of a stable component (T) and a random component (E). The observed score is the number of points an individual receives on a given test. This score fluctuates based on the amount of random error present. Therefore, examiners cannot assume that the observed score is an accurate representation of an individual's true abilities or the true score (Harvill, 1991).

Error score. The term E, defined as random error (Lord & Novick, 1968; Pedhazur & Schmelkin, 1991), is a random variable that does not include a stable or systematic component. The following assumptions define the nature of *E*. Across test takers, scores on *E* are uncorrelated with scores on *T* ($r_{et} = 0$, Equation 2.7.1b, Lord & Novick, 1968, p. 36). Across test takers, scores on *E* from an administration of parallel forms of a test are uncorrelated ($r_{e1e2} = 0$; Equation 2.7.1d, Lord & Novick, 1968, p. 36). Finally, across administrations of the same test to a given test taker, the expected value for *E* is zero ($E(E) = 0$; Equation 2.4.2, Lord & Novick, 1968, p. 31). CTT accepts the concept of error such that there will generally be random error present in measurement for reasons that include administration or procedural variations, environmental factors, instrumental limitations, or other factors. Depending on the nature of the random error, an individual's observed score can be affected either positively or negatively. For example, an individual may accidentally circle an incorrect answer on a question for which they knew the answer, an error which would result in a negative error score and an observed score lower than the true score. On the other hand, an individual may guess correctly on a question for which they did not know the answer, an error which would result in a positive error score and an observed score higher than the true score.

True score. The true score represents the part of the observed score that is free from random error of measurement (literally: $T = X - E$). Unlike the observed score, CTT assumes that the true score remains constant over time. *T* is the expected value of *X* across repeated measurements (i.e., $E(X) = T$; Equation 2.3.1, Lord & Novick, 1968, p. 30) as $X = T + E$ and the expected value of *E* is zero (thus, $E(X) = T + 0$). The true score is the theoretical value that represents an individual's score if no random error is present.

Systematic error in classical test theory. Systematic error occurs in the same manner in repeated measurements of a test taker and influences the mean of these repeated measurements as much as it influences each individual measure (Guion, 1965). A few examples of systematic error are test wiseness, rater leniency, or a faulty answer key; in all of the cases, scores in each instance will be affected in the same way.

After considering the particular distinctions between each component of the CTT equation, Guion (1965) proposed it may be helpful to rephrase the basic equation as:

$$X = s + e$$

Guion (1965) explained, "Now, instead of t (true measure), the equation considers s (systematic measure) to be a composite of a true measure and any constant error. In this revision of the equation, e represents only that residual error which is random and unpredictable" (p. 29).

Although it may be appealing to have both random and systematic measurement errors contained in the E component (leaving T as the errorless term within the classical test theory model), the assumptions of classical test theory do not allow for such a structure. It is crucial to understand that the names of the terms do not define their characteristics; the corresponding assumptions define these characteristics.

Reliability. The consistency and dependability of test scores is important for one to make meaningful inferences about those scores (Harvill, 1991). The CTT model of a test score (observed score equals the sum of the true score and error score) can also be expressed in terms of variance: the variance of the observed score equals the variance of the true score plus the variance of the error scores (Gulliksen, 1950).

$$S_X^2 = S_T^2 + S_E^2$$

From this variance expression, CTT defines reliability as the ratio of true score variance to observed score variance.

$$r_{XX} = \frac{S_T^2}{S_X^2}$$

The reliability coefficient for a given test reveals the extent to which the observed score variance is due to true score variance. True score and observed score variances would be equal in the case of perfect reliability (+1.00). The reliability coefficient indicates the percent of observed score variance that is not random error; lower levels of random error will lead to a higher reliability coefficient. The reliability of a test is estimated from obtained test scores from a group of examinees and can provide examiners with a good indication of whether measurement errors may be present or absent for the given group. However, reliability does not allow for the assessment of individual scores (Harvill, 1991). Reliability coefficients can be estimated through a variety of methods such as alternate forms, split-half, test-retest, and interrater reliability methods. Each of these procedures provide examiners with a value that estimates how free the test or measurement is from random error. Once the reliability coefficient is determined, further steps can be taken to estimate the role random error plays in an individual score.

Confidence Intervals and True Scores

As discussed, individual obtained or observed scores are collected upon completion of a test or measurement. Examiners know the observed score, but do not know the error score or true score. The observed score is insufficient because it does not necessarily represent an individual's true ability or true score due to the presence of random error. Confidence intervals are utilized to assist in determining the likely range of values of an individual's true score. The theoretical true score rather than the observed score should be considered in applied personnel activities. The addition of confidence intervals to observed scores is an effective way to report test scores to examinees or other interested persons and allows the unreliability of test scores to be expressed in a nontechnical way (Harvill, 1991).

Confidence intervals were first introduced to statistical hypothesis testing by Neyman (1937) and play a prominent role within CTT. Using an upper and lower limit on the score scale, confidence intervals produce a range of possible test scores within which an individual's true score is likely to exist (Harvill, 1991). These intervals allow for probabilistic statements about a true score. As with significance tests, confidence intervals are based on the standard error of the statistic. The application of confidence intervals to observed scores is complicated in that confidence intervals in classical test theory can be structured two different ways which call for two different standard error formulations, the standard error of measurement (SEM) and the standard error of estimate (SEE).

Standard Error of Measurement

The standard error of measurement is the standard deviation of errors of measurement that is associated with the test scores for a specified group of test taskers (AERA, APA, & NCME, 2014). The SEM equation provides the average magnitude of random error on a test for a given true score. The SEM is calculated by subtracting the reliability of the test from one, taking the square root of that difference, and multiplying the square root value by the standard deviation of the test scores (Dudek, 1979; Harvill, 1991).

$$SEM = S_X\sqrt{1 - r_{XX}}$$

A test's reliability is directly related to the SEM, which emphasizes the importance of obtaining a sound estimate of the reliability coefficient. If the reliability of a test equals zero, the SEM will equal the standard deviation of the observed test scores. If a test reliability is perfect (equaling one) the SEM will be zero. Higher reliability means less random error is present in individual observed scores. The SEM and the reliability coefficient each provide valuable information, but the SEM allows one to make statements regarding error at the individual score level, whereas the reliability coefficient is an index of the error present in the test as a whole (Ghiselli, Campbell, & Zedeck, 1981).

To use the SEM to address error at the individual score level, a confidence interval is formed. Typically, 95% confidence intervals are used, although 99% or 68% intervals are also employed. As Cascio, Outtz, Zedeck, and Goldstein (1991) stated, the "SEM 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). In other words, 5% of the observed scores will deviate from their true score by more than 1.96 SEMs.

Because the SEM indicates the standard deviation of observed scores if the true score is held constant, any use of this equation to form a confidence interval around a given observed score to determine the likely location of a true score is a misapplication of the SEM equation (Dudek, 1979). Dudek noted that although textbooks have called attention to this misuse of the SEM for years (e.g., Guilford, 1954; Lord & Novick, 1968; Nunnally, 1978), the equation continues to be misused. Dudek (1979) highlighted this misconception in order to guide and inform readers of the appropriate method, the standard error of estimate.

Standard Error of Estimate

As previously discussed, the SEM serves as an estimate of the variability expected for observed scores when the true score is held constant. Use of the SEM to set confidence intervals to locate true scores is in error (Dudek, 1979). Setting confidence intervals in search of true scores requires an index of error when the observed score is held constant (Dudek, 1979). This equation necessary for this application is given by Lord and Novick (1968) and is referred to as the standard error of estimate (SEE). The equation for SEE is similar to the SEM equation as it contains all of the same components in a slightly different arrangement.

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

Compared to the SEM, the SEE is smaller by a factor of the square root of the reliability ($SEE = SEM\sqrt{r_{XX}}$).

In addition to his arguments regarding the correct form of standard error for confidence intervals around an observed score, Dudek (1979) also argued that the interval should be based on an adjusted version of the observed score. This adjustment is a regression to the mean adjustment in which the observed score is moved closer to the mean (i.e., made less extreme). This adjustment is needed because extreme observed scores are often extreme due to the presence of large (in the same direction) E scores. That is, in these cases E is often very different from the expected value of E (i.e., 0), inflating the observed score. Thus, when applying confidence intervals in search of an individual's true score, one should not simply use the observed score value but the observed score value regressed toward the mean. This adjusted value is calculated by subtracting the mean of observed scores from the observed score, multiplying this difference by the test reliability, and adding the product to the observed score mean.

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

In summary, the suggested applications necessary for an accurate calculation of confidence intervals around an observed score with the goal of locating the true score includes the use of both the SEE equation (in place of the SEM equation) and the observed score value regressed toward the mean (Dudek, 1979). Proper procedures for locating the true score given an observed score are important as applications of these procedures extend beyond academic psychometric applications to applied decision making procedures (e.g., statistical banding).

The Present Study

The present study employed Monte Carlo analytic techniques to explore the accuracy of the various ways to compute a confidence interval to locate a true score given an observed score. A Monte Carlo analysis is a statistical technique that generates large datasets to test statistical models and procedures. Monte Carlo analyses have the benefit of allowing the researcher to explore the effectiveness of statistical methods in a variety of conditions with large sample sizes, thus serving as an appropriate means for evaluating this study's proposed research question.

I expected that a SEE based 95% confidence interval formed around an observed score regressed to the mean will exhibit the best performance; this interval will include the true score with the desired accuracy (i.e., 95%) and will evenly balance true scores outside of the interval (i.e., overestimates equal underestimates). Additionally, the SEE interval has the advantage of hitting both of these goals with a narrower width than a SEM interval.

Method

Design

A Monte Carlo analysis was executed to assess the accuracy of both the SEM and SEE equations in conjunction with the presence or absence of a regression to the mean adjustment to the observed score. Thus, there were four types of confidence intervals computed within each condition: a SEM based confidence interval with and without a regression to the mean adjustment to the observed score and a SEE based confidence interval with and without a regression to the mean adjustment to the observed score.

Only one variable, reliability, was manipulated for this study. The accuracy of the four methods for computing the confidence interval was evaluated at nine levels of reliability ranging from .1 to .9 (in .1 increments).

Sample

I used the random number generator function (normal distribution) of SAS (SAS Institute, 2013) to generate a true score and error score for each case. The use of randomly generated variables satisfies the CTT assumption of uncorrelated true scores and error scores as random variables are uncorrelated with other variables. Observed scores were computed as the sum of the true and error scores. One million cases were generated for each condition. Results were averaged across these one million trials.

Dependent Variables

Because this is a Monte Carlo design and because true scores and error scores (as well as the resultant observed scores) are known for each case, the success of the confidence interval at locating the true score could be assessed. For reliability levels ranging from .1 through .9 (in .1 increments), the accuracy of the 95% confidence intervals were assessed for the four confidence interval formulations.

$CI_{SEM\_Norm} = X$ +/- (1.96*SEM)

$CI_{SEM\_RTM} = X_{RTM}$ +/- (1.96*SEM)

$CI_{SEE\_Norm} = X$ +/- (1.96*SEE)

$CI_{SEE\_RTM} = X_{RTM}$ +/- (1.96*SEE)

If the confidence interval contained the true score, then the outcome was coded as successful for that confidence interval. Thus, the dependent variable is dichotomous with 1 representing a true score falling within the given confidence interval and 0 representing a true score falling outside of the interval. A 95% confidence interval functioning correctly will include the true score in 95% of the trials. The mean success of the confidence interval was computed across the one million trials.

In addition, for those cases for which the SEM based confidence interval without a regression to the mean adjustment (i.e., the most common form of this interval) failed to include the true score, we also assessed whether the true score was located between the population mean and the lower bound of the interval (meaning the interval was too extreme) or beyond the upper bound of the interval (meaning the interval was too conservative).

Results

The means for the four proposed confidence interval equations across the nine reliability conditions are listed in Table 1. Each mean represents the percentage of cases in which the confidence interval successfully captured the true score, with 95% accuracy being desired. The SEM based confidence interval with a regression toward the mean adjustment demonstrated the highest percentage, capturing the true score in 99% of the trials. If a confidence interval designed to capture the true score in 95% of the cases contains the desired value more than 95% of the time then the interval must be considered too wide. The least accurate equation is the SEE based confidence interval without a regression to the mean adjustment as it produced an average accuracy rate of only 78%. Appearing appropriately accurate are the SEM based confidence interval without a regression toward the mean adjustment and the SEE based confidence interval with a regression toward the mean adjustment, each with ideal average accuracy rates of 95%. Although these two equations are tied for accuracy, we must recognize that the SEM based confidence interval without a regression to the mean adjustment produces a wider (by factor of $1/\sqrt{r_{XX}}$) confidence interval than the SEE based confidence interval with a regression toward the mean adjustment. Therefore, given that the accuracy of the intervals is equal, the SEE based confidence interval with a regression toward the mean adjustment is more useful than the SEM based confidence interval without a regression toward the mean adjustment.

In short, the SEM based 95% confidence interval (without a regression to the mean adjustment) that has been traditionally used does capture the true score 95% of the time. However, this confidence interval is inefficient as it is wider than is necessary. The SEE based interval with a regression to the mean adjustment should be preferred.

Table 1

Proportion of Trials Where a Confidence Interval on the Observed Score Contained the True Score for Different Methods of Computing the Confidence Interval

Reliability (r_{XX})	SEM-NORM	SEM-RTM	SEE-NORM	SEE-RTM
	M	*M*	*M*	*M*
0.1	0.949	1.000	0.464	0.949
0.2	0.950	0.999	0.619	0.949
0.3	0.949	0.999	0.716	0.950
0.4	0.950	0.997	0.784	0.949
0.5	0.949	0.994	0.833	0.950
0.6	0.949	0.988	0.870	0.949
0.7	0.949	0.980	0.898	0.949
0.8	0.949	0.971	0.920	0.949
0.9	0.950	0.961	0.936	0.950
Total	0.949	0.988	0.782	0.949

Note. N = 1000000 for individual reliability conditions, N = 9000000 for total.

As an additional analysis of the remaining 5% of cases not captured by the SEM based confidence interval without a regression toward the mean adjustment, we examined where the confidence interval missed the true scores. Table 2 displays the number of cases and the percentage of these cases in which the true score was located between the population mean and the lower bound of the interval (meaning the confidence interval was too extreme) across nine conditions of reliability.

Table 2

Location of True Scores Outside of SEM-NORM CI

Reliability (r_{XX})	N cases not in CI	Percent beyond upper bound	Percent below lower bound
0.1	50,044	0.0%	100.0%
0.2	49,788	0.0%	100.0%
0.3	50,243	0.0%	100.0%
0.4	49,852	0.0%	100.0%
0.5	50,189	1.2%	98.8%
0.6	50,190	3.2%	96.8%
0.7	50,035	6.7%	93.3%
0.8	50,112	12.3%	87.7%
0.9	49,779	21.8%	78.2%

As shown in Table 2, the SEM based confidence interval without a regression toward the mean adjustment, in addition to being unnecessarily wide, is biased as regards the cases for which it fails to contain the true score. The data in Table 2 demonstrate that at all reliability levels the true score falls between the mean and the lower bound of the interval far more often that it falls beyond the upper bound. Thus, the SEM based confidence interval without a regression toward the mean adjustment was not only wider than the SEE based confidence interval with a regression toward the mean equation, but also missed the true scores in an uneven manner. Ideally, there should be an even balance of the missed true scores. Most of the time, the true score was not captured by the confidence interval due to the interval resting too far from the mean. This result explains why the regression toward the mean adjustment to the SEM equation allows the confidence interval to capture an ample amount of otherwise missed cases in the SEM equation without the regression toward the mean adjustment.

Discussion

Dudek (1979) argued that confidence intervals formed around observed scores for the purpose of locating true scores are in error when constructed with the SEM equation. The purpose of this study was to empirically test this claim and compare the accuracy and efficiency of various methods for computing the confidence interval around an observed score. I-O psychologists and other personnel decision-making professionals regularly use statistical tools such as confidence intervals to interpret test scores and guide organizational procedures such as selection, promotion, training, development, and performance management. Therefore, it is essential that the mathematical formulas calculating the confidence intervals are applied in the appropriate manner.

Results produced by this Monte Carlo simulation allowed for the simple rejection of the SEE based confidence interval without a regression toward the mean adjustment with the accuracy rates falling well below the 95% target. Both the 95% SEE based interval with a regression toward the mean adjustment and the 95% SEM based interval without a regression toward the mean adjustment appeared to perform as advertised as they capture the hypothetical true score 95% of the time. Although both equations demonstrated the same accuracy, the SEE based interval with a regression toward the mean adjustment has an advantage in efficiency as it produces a narrower interval than the SEM based confidence interval without a regression toward the mean adjustment. It is self-evident that a narrower interval is more useful in locating an individual's true score.

Additional analysis of the SEM based confidence interval without a regression toward the mean adjustment demonstrated a further inefficiency in that the calculated interval was located too far from the mean, resulting in missed true scores that were far more likely to be below the lower bound of the mean than beyond the upper bound.

Practical Implications and Future Research

Standard error of the difference. The commonly computed SEM value is traditionally used in a supplemental calculation known as the standard error of the difference (SED; Gulliksen, 1950). The SED provides a range of observed scores in which one cannot deem significantly different from one another because of the possible range of true scores (Cascio et al., 1991). Gasperson et al. (2013) conducted a study that specifically concerned the potential effects of calculating the SED with the SEM vs. the SEE and found substantial variations in banding-based selection decisions depending on whether the SED formula used the SEM value or SEE value. These observed variations (selection means, selections by race, and minority selection ratios) are a result of smaller bands produced when the SED formula is employed using the SEE value (Gasperson et al., 2013).

Statistical banding. Statistically based banding or test score banding is a technique within the field of psychometrics that uses the SED to guide employment decisions. The SED allows one to create a range of observed scores that are deemed equivalent based on the assumption that solitary observed scores are considerably unreliable (Gasperson et al., 2013). The concept of statistical banding began to emerge in the mid-1980s and serves as an alternative to strict top-down selection (Sproule, 1984). Specifically, this additional method was introduced in order to address concerns of adverse impact while also minimizing the loss of utility (Cascio et al., 1991).

Given the proposed inaccuracy of the SEM equation per Dudek (1979), Gasperson et al. (2013) conducted a study to explore the correct way in which bands should be established. Differences in employment decisions were assessed when bands were created using the SEM based SED versus the SEE based SED. Although only two data sets were examined, Gasperson et al. (2013) noted variations in band sizes and selection decisions while comparing the two SED procedures, supporting the modifications proposed by Dudek (1979). When the SEM based SED was used, larger bands were produced which led to at least one employee who was erroneously accepted and another who was erroneously rejected (Gasperson et al., 2013). Given the gravity of employment decisions and the potential legal implications of such decisions, practitioners must ensure they are referring to the most accurate formula and method in guiding their personnel practices.

Diversity and inclusion. As diversity and inclusion currently serve as popular topics among organizations, pressure on proper selection strategies that work to minimize adverse impact will only continue to increase. Racial, gender, disability, and veteran representation within an organization depends upon initial selection procedures. Gasperson et al. (2013) presented evidence supporting this concern in the form of differences in band sizes and selection decisions when comparing the use of the SEM based SED and the SEE based SED. These findings indicated the need for additional research in this area such as the simple exploration of additional data sets to ensure professionals are using the most statistically sound method.

Organizational resources. In addition to ethical and legal concerns, the proper application of the SEE equation can also result in the preservation of organizational resources. By accurately estimating individual true scores, thus increasing precision in selecting the most capable/qualified candidates for employment or promotion, an organization could potentially save both time and money. This increased precision can in turn have a positive impact on overall organizational success as the most fitting applicants with the highest potential for productivity and development are selected.

Conclusion

As indicated through the findings in this analysis, the comparison of SEM and SEE equations (with and without mean adjustments) are worthwhile. Differences among the confidence interval ranges and location of uncaptured true scores should continue to be noted, along with further examination of differences in SED calculations and potential banding variations. Applying these considerations to data sets varying in both size, reliability, and type (real world data vs. simulated data) would be beneficial in this area of study.

References

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington D.C. American Psychological Association.

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.

Guilford, J. P. (1954). *Psychometric methods.* New York, NY: McGraw-Hill.

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.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236*, 333-380.

Nunnally, J. C. (1978). *Psychometric theory.* New York, NY: McGraw-Hill.

Pedhazur, E. J., & Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

SAS Institute Inc. 2013. *SAS® 9.4 Statements: Reference*. Cary, NC: SAS Institute Inc.

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

Sproule, C. F. (1984). Should personnel selection tests be used on a pass-fail, grouping, or ranking basis? *Public Personnel Management, 13*, 375-394.