Order Number 1339233

Standard errors of measurement, confidence intervals, and the distribution of error for the observed score curve

Tataryn, Douglas Joseph, M.A.

The University of Arizona, 1989

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

STANDARD ERRORS OF MEASUREMENT, CONFIDENCE INTERVALS, AND THE DISTRIBUTION OF ERROR FOR THE OBSERVED SCORE CURVE

by

Douglas Joseph Tataryn

A Thesis Submitted to the Faculty of the

DEPARTMENT OF PSYCHOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF ARTS

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 8 9

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Doug Tatarzyn

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

Lee Sechrest
Lee Sechrest
Professor of Psychology

Dec 12/89
Date

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS 4

ABSTRACT 5

CHAPTER

1. THE STANDARD ERROR OF MEASUREMENT 6

2. THE DISTRIBUTION OF ERROR ABOUT THE OBSERVED CURVE . . . 11

3. EXAMINING THE PROPOSALS FOR THE USE OF SEM 18

4. ALTERNATIVE FORMULATIONS OF THE SEM 22

5. REASSESSING THE UTILITY OF THE SEM FOR OBSERVED SCORES . 24

6. THE DISTRIBUTION OF OBSERVED SCORES AROUND THE E(T) SCORE 25

7. SUMMARY 27

REFERENCES 28

LIST OF ILLUSTRATIONS

FIGURE 1: Error variance as a function of standardized deviation score and reliability . . . .

ABSTRACT

This paper reviews the basic literature on the suggested applications of the standard error of measurement (SEM), and points out that there are discrepancies in its suggested application. In the process of determining the efficacy and appropriateness of each of the proposals, a formula to determine the distribution of error for the observed curve is derived. The final recommendation, which is congruent with Cronbach, Gleser, Nanda & Rajaratnam's (1972) recommendations, is to not use the SEM to create confidence intervals around the observed score: The predicted true score and the standard error of the prediction are better suited (non-biased and more efficient) for the task of estimating a confidence interval which will contain an individual's true score. Finally, the distribution of future observed scores around the expected true score is derived.

CHAPTER 1

THE STANDARD ERROR OF MEASUREMENT

In classical test theory, the error associated with the measurement of an individual's true score is assumed to be randomly and normally distributed with a mean of zero. Unfortunately, the repeated testing necessary to estimate the error of measurement for an individual is not generally practical. A pragmatic compromise between the theoretical definition of measurement error and the realities of the testing world is to assume that all individuals have approximately equal error distributions associated with the measurement of their true score, and to use the average error of measurement as an estimate of the error distribution for each individual. The formula for the (average) SEM is;

$$SEM = \sqrt{\sigma^2 x\ (1 - r)} \qquad (1)$$

where $\sigma^2 x$ is the variance of the observed scores, and r is the reliability associated with the instrument on which the observed scores are based. Simply put, the SEM is the square root of the proportion of observed variance which is not attributable to real differences between people. The SEM, in conjunction with the assumption of a normal distribution of error, can be used to construct confidence intervals (CI) which, upon repeated testing, will contain a known proportion of observed scores.

A problem with a direct application of the SEM is that it delineates the distribution of observed scores around the true score. Unfortunately, the true score, being defined as the average of an infinite number of parallel tests, is never known. Given this circumstance, the actual utility and application of the SEM is no longer obvious. This is evidenced by the fact that different writers in the area of psychometrics have conflicting opinions regarding the proper use of the SEM.

Gulliksen (1950) uses the SEM to define the "reasonable limits" (three SEM on each side of an observed score) in which an individual's true score will likely lie. This concept is based on the fact that 99% of the observed scores will not vary more than three SEM from the true score, and hence, any true score is likely to be within three SEM of a corresponding observed score. He does not define reasonable limits as a 99% confidence interval (which would necessitate a normal distribution of true scores around the observed score), nor does he assign a probability to the interval containing the true score. Gulliksen's logic is also endorsed by Anastasi (1982).

While Crocker and Algina (1986) do not explicitly cite Gulliksen, they use a similar logic to generalize beyond his "reasonable limits" to a formulation of a type of confidence interval. Their argument is as follows. If a person is tested

100 times, and for example, ± 1 SEM confidence bands are drawn around each of the observations, 68% of them would contain the true score. Hence,

> "[o]n a single testing, there is a 68% chance that we will obtain one of the observed scores that lies [within ± 1 SEM of the true score]." (p. 122)

The same logic follows for any CI, the probability of the interval containing the true score being a function of the probability density of observed scores around the true score. Crocker and Algina's procedure generates a statement of the probability that a given interval will contain the true score. I will refer to these as "confidence statements".

Lord (Lord & Novick, 1968) states that a person's true score probably lies in the interval defined by ± 2 SEM around the observed score. He then qualifies this.

> ...one <u>cannot</u> have confidence in the correctness of such an inference for any single, nonrandomly chosen individual... In particular, it should be understood that <u>confidence intervals</u> obtained in this way are not valid if we become interested in... [the person] ...as a result of seeing their test scores. (p. 160, second emphasis mine, parenthetical material mine)

Lord is emphasizing that any <u>randomly</u> selected observed score is an unbiased estimate of the true score and, hence, can be substituted for it and used to define confidence intervals. If on the other hand, the observed score is not randomly selected, but selected for some reason, say for where it falls in the total distribution, the observed score is a biased estimate of the true score and not suitable for defining a CI.

Cronbach's (1970) statement regarding the appropriate use of SEM is less favorable. After discussing the formal meaning of the SEM he states;

> Some interpreters reverse this reasoning, arguing that if someone's observed score is, say, 47, his universe [true] score then probably lies in the range 47 ± 2 s.e.m. If the s.e.m. is 3, they would say that 41-53 is a "95 per cent confidence interval" for the person's universe score. This technique is likely to give false conclusions - particularly for persons with extreme observed scores - and cannot be recommended. (p. 164, parenthetical material mine).

Each proposal for the functional use of the SEM; Gulliksen's reasonable limits, Crocker and Algina's confidence statements, and Lord's confidence interval, appear to be cogent within the contexts they have discussed. Lord and Cronbach, however, have raised issues that should bear in the consideration of any of the suggested procedures. The first issue deals with the selection procedure (random vs non-random) of the observed score of interest, and the second deals with the extremeness of the score relative to the rest of the distribution.

Neither Gulliksen nor Crocker and Algina address the issue of observed score selection nor the issue of extremeness. In not doing so, one must assume they are endorsing their procedures for any observed score. Lord endorses his procedure only for randomly selected observed scores. His judgment, however, seems to be at odds with Cronbach's. For example, suppose that an observed score is chosen randomly, and, quite by chance, it is extreme in the sense that Cronbach means. According to Lord, a

CI can be constructed around the score, and the interval used in good conscience to make a statement regarding the likelihood of it containing the true score. Yet Cronbach states that since the score is extreme, the CI would be inappropriate as it would be erroneous. How are these two opinions reconciled, and just how appropriate are the other suggested procedures when considered in the context of selection procedure and relative deviancy? In order to answer this question, it is useful to remember that the SEM describes the error distribution around true scores, but that it is being used with observed scores. It would seem that exploring the distribution of error for the observed curve may help to form a context in which to evaluate both the different procedures and the cautions regarding their use.

CHAPTER 2

THE DISTRIBUTION OF ERROR ABOUT THE OBSERVED CURVE

One of the basic tenets of classical test theory is that true scores and error scores are uncorrelated. That is, the magnitude of error associated with the measurement of a trait is equivalent across all points of the true score distribution. Observed scores on the other hand, being defined as the sum of both true scores and error components, are necessarily correlated not only with the true scores, but with the error components as well. The correlation between observed scores and true scores and observed scores and error, is;

$$r_{tx} = \sqrt{r} \quad (2)$$

$$r_{ex} = \sqrt{1 - r} \quad (3)$$

(See Gulliksen, 1950, pp. 33-35 for the derivations)

where r_{tx} is the correlation of the true scores with the observed score, r_{ex} is the correlation of the error components with the observed score, and $\underline{r}$ is the reliability of the test.

These correlations imply that, on average, a certain proportion of an observed score will be attributable to the true score and a certain proportion to the error component. Continuing with this logic, an observed score of x standard deviations above or below the mean on a test with reliability $\underline{r}$ will, on average, have r proportion of that score contributed from the true score, and 1 - r contributed from an error component. For example, a score of 2 standard deviations above the mean of a test with reliability of .7, will have an average true score component of 2 * .7 = 1.4 standard deviations and an average error component of 2 * (1 - r) = 2 * .3 = .6 standard deviations. Hence the an observed score, in the context of the rest of the distribution, can be used to estimate the true score for which it is most likely a fallible observation. This predicted value, by way of definition in the regression formula, is uncorrelated with the residuals, or error, components. The general formula for the expected or predicted true score is;

$$E(T) = \bar{X} + r\ (X - \bar{X}) \qquad (4)$$

where E(T) is the expected true score, $\bar{X}$ is the mean of the distribution, r is the reliability of the test, and X is the raw observed score. The standard error of the estimate (SEE) of the E(T) is;

$$SEE = \sigma_X \sqrt{r\ (1 - r)} \qquad (5)$$

(as derived in Gulliksen, 1950)

To illustrate, suppose an individual obtains a score of 80 on a test with a reliability of .7, mean of 50, and a standard deviation of 10. To estimate that individual's most likely true score, we calculate the product of the observed score deviation (from the mean) and the reliability; 30 * .7 = 21, and add this to the mean; 50 + 21 = 71. Thus though the individual scored 80 on the test, his most likely true score, given the inaccuracies of measurement associated with that test, is 71. This estimate of his true score is the most likely of many possible values. The distribution of the other possible values is normally distributed with a standard deviation of 10 * the square root of (.7 * (1 - .7)) = 10 * .46 = 4.6 units. Thus, a confidence interval can be set up around the estimated true score that will contain the real true score a known percentage of the time. A 95% confidence interval would be 71 ± 1.96 * 4.6 = 71 ± 9 = 62 to 80; there is a 95% probability that the interval 62 to 80 contains the individual's true score.

Having calculated the expected true score component of the observed score, we can also calculate the expected or average error component. This is simply the observed score minus the expected true score;

$$E(E) = X - E(T) \qquad (6)$$

which for our example would be 80 - 71 = 9 units.

Since the distribution of error around the expected true score is normally and symmetrically distributed about the mean E(T), and the observed score differs from the mean of that distribution by a known constant, E(E), we can enter this constant into our equation for calculating a variance and determine the exact variance of errors associated with the observed score.

Variance is defined as $\frac{\Sigma(\overline{X} - X)^2}{N}$

The variance calculated E(E) units from the mean (which for generality and simplicity in the derivation I will refer to as C, any constant) would be:

$$\frac{\Sigma(X - (\overline{X} + C))^2}{N}$$

Expanding the top portion out;

$$\Sigma (X^2 - 2X(\overline{X} + C) + (\overline{X} + C)^2)$$

$$\Sigma (X^2 - 2X\overline{X} - 2XC + \overline{X}^2 + 2\overline{X}C + C^2)$$

Separating the components with C in them;

$$\Sigma ((X^2 - 2X\overline{X} + \overline{X}^2) + (C^2 + 2\overline{X}C - 2XC))$$

$$\Sigma ((X^2 - 2X\overline{X} + \overline{X}^2) + (C^2 + 2C (\overline{X} - X))$$

Reducing the first half of the equation and distributing the summation sign;

$$\Sigma(\overline{X} - X)^2 + \Sigma C^2 + 2C\,\Sigma(\overline{X} - X)$$

Since the sum of the observation from the mean is zero by definition of the mean, the last term drops out, and since the summation over N of a constant is simply N times the constant, this equation becomes;

$$\Sigma(\overline{X} - X)^2 + NC^2$$

Finally, replacing the terms back into the original equation, we have;

$$\frac{\Sigma(\overline{X} - X)^2 + NC^2}{N} = \frac{\Sigma(\overline{X} - X)^2}{N} + C^2$$

Hence, the variance of a distribution calculated from a point other than the mean is simply the square of the distance of the mean from the new calculation point (C) plus the variance of the distribution. In calculating the variance of error for points along the observed curve, the distance (C) is the expected error, E(E), for any given observed score, and the variance of the distribution of expected scores is SEE squared. Substituting these terms into the derived equation, we have;

$$\sigma^2 ex = \sigma^2_X \ (r \ (1 - r)) + E(E)^2 \tag{7}$$

where $\sigma^2 ex$ is the error variance for score X on the observed curve, r is the reliability, and $E(E)^2$ is the square of the expected error for the score on the observed distribution. Note that the error variance for the observed curve is a curvilinear function, dependent on the variance of the observed scores, the reliability of those observed scores, and the distance from the mean of the group. Figure 1 presents the error variance for the standardized observed score distribution (mean of zero, unit variance), as a function of the observed scores' deviations from the mean, for different reliabilities. We can see that the error variance increases exponentially as a function of the distance from the mean, and that the magnitude of the error increases substantially more quickly for instruments of lower reliability. Examining the graph in detail, for example, for an instrument with a reliability of .5, a deviation score of two units from the

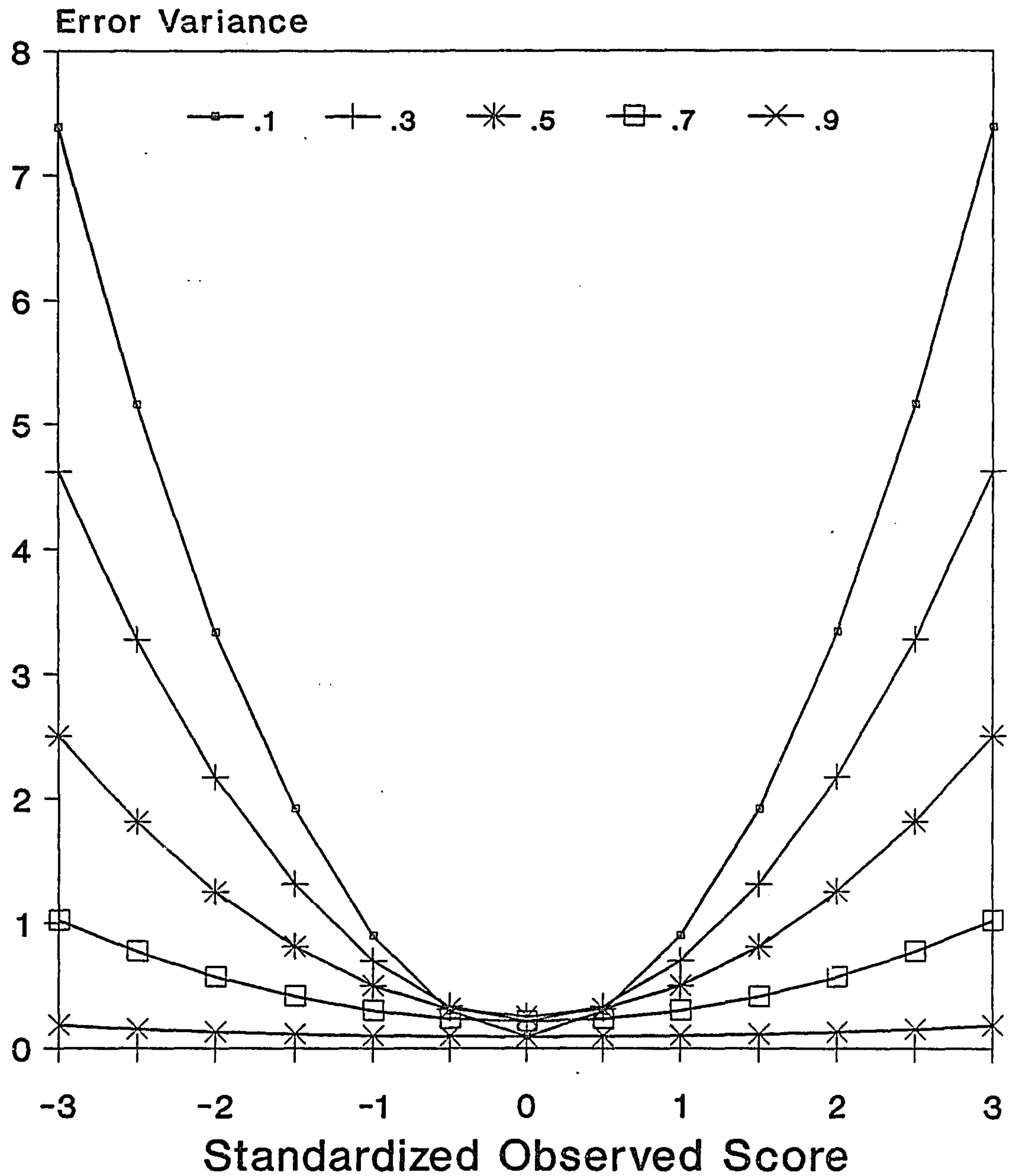

Figure 1: Error variance as a function of standardized deviation score and reliability

mean has an error variance of 1.25, two and a half times that estimated by the SEM^2 (1 - r = error variance of .5). In general, the error variance is larger than the SEM^2 for people with observed scores whose magnitude is greater than unity, equal to SEM^2 for scores of -1 and +1, and is smaller than SEM^2 between -1 and +1, converging on SEE^2 at the mean. Without going into the calculations, if one multiplies the variance associated with each deviation score by the probability density of that score, the total average error variance is SEM^2. This means that while the SEM is only accurate for points on the observed curve of ± 1, it is the expected or average error for the whole observed curve.

CHAPTER 3

EXAMINING THE PROPOSALS FOR THE USE OF SEM

Using the information on the exact distribution of error variance for the observed curve, we now reexamine the various proposals for the use of the SEM and confidence intervals based on it. Lord's use of a confidence interval was contingent on the random selection of the observed score. However, as the present discussion has made clear, given the assumptions of classical test theory, random selection only ensures that the expectation of the observed score will be the true score and that the expectation of the standard errors will be the SEM. For any individual observed score, both statistics will err from their actual parameters, with the observed score consistently overestimating the magnitude of the true score (except for scores at the mean), and the SEM over- or under-estimating the actual error associated an observed score except for observed scores of ± 1. What has become obvious, and which agrees with Cronbach's (1970; et al, 1972) reasoning, is that the observed score must be placed in the context of the distribution for which it is a member: regardless of selection procedure, the simple application of the SEM to create confidence intervals for an observed score will always be erroneous[1].

Let us use for example, a standardized observed score of 3 (an "extreme" score), on a test with a reliability of .5. The SEM is the square root of .5, which is .707; therefore a 95% confidence interval for containing the true score is 3 ± 1.4, which is roughly 1.6 to 4.4. We now calculate the expected true score to be 3 * .5 = 1.5, with an SEE equal to the square root of (.5 * (1 - .5) = .5. The lower bound of the confidence interval based on the observed score and the SEM is 1.6, which is (1.6 - 1.5)/.5 = .2 SEE deviations from the expected true score. The proportion of true scores actually falling above .2 SEE deviations is 42%. Thus the 95% confidence interval, based on the SEM, and centered at the observed score, will actually contain only 42% of the true scores.

Gulliksen, who set his reasonable limits to contain a true score at ± 3 SEM, is fairly safe, as this interval will encompass most of the distribution of true scores except at the most extreme values. More specifically, a 99% probability interval ranges from 3 ± 3 * .707 = 3 ± 2.1 = .9 to 5.1. The proportion of true scores to right of .9, or (.9 - 1.5)/.5 = -.6/.5 = -1.2

---

1 While the observed score is a non-biased estimate of the true score when it is equal to 0, the SEM overestimates the error of measurement at that point. When the error of measurement is accurate (at ± 1) the observed score is a biased estimate of true score. Hence the SEM used with an observed score to create confidence intervals is always misleading.

SEE is roughly 89%. Since Gulliksen never associated an exact probability to his reasonable limits, the question whether 89% is a reasonable accuracy for his interval is one of judgement. Aside from the fact that the CI is a rather wide one, I think few people would argue that almost 90% accuracy under these conditions is not too misleading an estimate.

For Crocker and Algina, who advocate the use of any range, the results become more inaccurate and misleading as the designated CI get smaller or the score more extreme. Using their CI of 68% on the same example: the lower bound of the CI is 3 - .7 = 2.3. This score of 2.3 is (2.3 - 1.5)/.5 = .8/.5 = 1.6 SEE from the mean of the expected true score. Hence the confidence statement that they present as being right 68% of the time will only be right 5% of the time.

It should be noted that because of the inherent asymmetry of true scores around the observed scores, the errors of omission made by the three procedures for estimating true score ranges are not symmetrical around the confidence interval. Almost all of the true scores not contained within a given CI interval are on the side closest to the overall mean of the distribution, and almost none on the side of the confidence interval farthest from the mean. Additionally, for scores between ± 1 on a standardized instrument, the CI presented by the three authors will actually <u>under</u>estimate the number of cases falling within the bounds. For example, for a score at the mean, with a reliability of .5, a 68%

CI lies between 0 ± .7, or -.7 to .7 units. Yet these parameters, relative to the SEE, are actually .7/.5 = 1.4 units. Thus the interval will actually contain 84% of the true scores, or be right 84% of the time, not the 68% estimated by the SEM procedure. By overestimating for some observed scores, and underestimating for other scores, the CI will, on average, contain the percentage of correct estimates it is supposed to. Unfortunately, while the expected value of an infinite number of CI is an unbiased estimate, the are no cases for which it is accurate for any particular observed score. (See footnote 1.)

CHAPTER 4

ALTERNATIVE FORMULATIONS OF THE SEM

One alternative to the SEM is to use the error variance associated with a given score on the distribution of observed scores to create a more appropriate standard error term. This error term will be called the standard error of measurement for an observed score (SEMobs). Using the situation from our prior examples, we can calculate the error variance for a score 3 standard deviations above the mean;

$$
\begin{aligned}
SEMobs^2 &= .5 * .5 + (3 * .5)^2 \\
&= .25 + 1.5^2 \\
&= .25 + 2.25 \\
&= 2.50
\end{aligned}
$$

The SEMobs would be the square root of 2.5, which is approximately 1.6 units. For Gulliksen's reasonable limits formulation, three SEMobs on either side of the observed score would be 3 ± 4.8 = -1.8 to 7.8. Since the -1.8 is (-1.8 - 1.5)/.5 = -3.2/.5 = -6.4 SEE units below the expected true score, and 7.8 an even larger number of SEE deviation units above, this interval would actually contain more than 99.9% of the true scores.

Lord's 95% confidence interval, using the SEMobs, would range from 3 ± 2*1.6 = 3 ± 3.2, which is -.2 to 6.2. This translates into a SEE deviation scores of (-2 - 1.5)/.5 = 1.7/.5

= 3.2 and (6.2 - 1.5)/.5 = 4.7/.5 = 9.2. Hence the 95% CI, based on the SEMobs is also very liberal in its bounds, containing the true scores over 99% of the time.

For the Crocker and Algina procedure, a 68% confidence statement, would be based on an interval from 3 ± 1.6 = 1.4 to 4.6 units. This range translates into SEE units of (1.4 - 1.5)/.5 = -.1/.5 = -.2 and (4.6 - 1.5)/.5 = 3.1/.5 = 6.2 units, and would contain the true scores 69% of the time, almost the probability that it is purported. Note however, that as the CI decreases, CIs based on the SEMobs will still become increasingly erroneous and biased.

In their favor, CIs based on SEMobs become smaller and less biased for deviation scores between ± 1, converging on the expected true score and its distribution at the mean of the observed score distribution. For the extreme scores, however, which the proposed SEM procedures are most inaccurate for, the SEMobs tend to produce such large confidence ranges that they tend to be of little practical utility. Another problem is that while SEMobs take into consideration the correct error variance associated with an observed score, CIs based on them are still symmetrical about the observed score, and asymmetrical with respect to the actual expected true score distribution. Thus using either the SEM or SEMobs in the context of observed scores has a inherent limitation: they become increasingly biased as a function of the distance from the mean of the observed scores.

CHAPTER 5

REASSESSING THE UTILITY OF THE SEM FOR OBSERVED SCORES

In review, the SEM is a measure of the error associated with the measurement of an individual's true ability. Theoretically it can be used to place confidence intervals around a true score, inside of which a predictable proportion of observed scores will fall. In attempting to make use of the SEM with observed scores, several problems must be considered. First is the fact that the error variance for a given point on the observed score distribution varies exponentially with the deviation distance from the mean of the distribution. The second, but related problem, is that observed scores are biased estimators of underlying true scores: a linear function of deviation distance from the mean. The reviewed proposals for the use of the SEM to create true score confidence intervals around observed scores fail to incorporate these two facts, and hence produce inaccurate, misleading intervals. On the other hand, by explicitly using this information to produce a standard regression equation, the expected true score and standard error of estimating the true score produce far superior estimates of the person's true score, ones that are non-biased, more efficient, and more precise than those based on observed scores and the SEM.

CHAPTER 6

THE DISTRIBUTION OF OBSERVED SCORES AROUND THE E(T) SCORE

The original derivation of the SEM was in the context of describing the distribution of possible future observed scores around a known true score value. Having examined the exact nature of expected true scores and its distribution, and also knowing the SEM, we are now in a position to readdress this original conceptualization. It can be thought of in the following way: The expected true score is the mean value of a normally distributed density of possible true scores. The standard error of measurement associated with the measurement of any true score is the SEM. The distribution of the measured values (observed scores) will be the distribution of true scores in conjunction with the error distribution. Since the error and the possible true scores are independent, the total expected variance of the distribution of observed scores around the expected true scores is simply the sum of the two distributions; $SEE^2 + SEM^2$. The square root of this variance is the standard error of observed scores around the expected true score, which is;

$$= \sigma^2_X \,(\, r\,(\,1 - r\,) + \sigma_2 X\,(\,1 - r\,))$$

$$= \sigma^2_X \,(\, r\,(1 - r) + (\,1 - r\,))$$

$$= \sigma^2_X \,(\, r - r^2 + 1 - r\,)$$

$$= \sigma^2_X \,(\,1 - r^2\,)$$

$$= \sigma_X \sqrt{1 - r^2} \qquad (8)$$

Hence the standard error of observed scores around the expected true score is the standard deviation of the observed scores multiplied by the square root of the difference between unity and the reliability squared. This value, as is seen in the fact that it is the sum of the SEM and the SEE, will always be larger than either of them. A reader familiar with regression techniques will recognize this equation as the error associated with predicting one fallible measure with one or more other fallible measures, which is in essence, what this procedure is doing.

CHAPTER 7

SUMMARY

This paper has outlined several existing proposals for utilizing the SEM to produce confidence intervals or statements regarding the likelihood of a given range containing an individual's true score. After considering the various proposals in the context of the distribution of error for the observed score curve, it was shown that the expected true score, with a SEE which is smaller than the SEM, is non-biased and precise, and hence a more appropriate estimator of such confidence intervals. This information was then used to derive the distribution of future observed scores around the expected true score value.

REFERENCES

Anastasi, A. (1982). Psychological Testing. MacMillan Publishing, Co., NY.

Crocker and Algina (1986). Introduction to Modern and Classical Test Theory. Holt, Rinehart & Winston, NY.

Cronbach, L.J. (1970). The Essentials of Psychological Testing. Harper & Row, NY.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). The Dependability of Behavioral Measures. Wiley, NY.

Gulliksen, H. (1950). Theory of Mental Tests. Wiley, NY.

Lord, F.M. and Novick, M.R. (1968) Statistical Theories of Mental Test Scores. Addison-Wesley, Mass.