An NCME Instructional Module on

Standard Error of Measurement

Leo M. Harvill, *East Tennessee State University*

The standard error of measurement (SEM) is the standard deviation of errors of measurement that are associated with test scores from a particular group of examinees. When used to calculate confidence bands around obtained test scores, it can be helpful in expressing the unreliability of individual test scores in an understandable way. Score bands can also be used to interpret intraindividual and interindividual score differences. Interpreters should be wary of over-interpretation when using approximations for correctly calculated score bands. It is recommended that SEMs at various score levels be used in calculating score bands rather than a single SEM value.

What is an error of measurement? What is an examinee's true score on a test and how is it different from the actual or obtained score? What is the standard error of measurement (*SEM*)? Is the *SEM* a characteristic of the test, an examinee's test score, or both? Does a particular test have a single *SEM* or different *SEM*s for different score levels? Is a score band estimating an examinee's true score symmetrical around the examinee's obtained score? If an examinee obtained a score of 40 on a test and the *SEM* was reported as 3 for that test, is it correct to state that "the chances are two out of three that the true score for the examinee is between 37 and 43?"

These are the kinds of questions which will be answered in this module. The *SEM* is a determination of the amount of variation or spread in the measurement errors for a test. A measurement error is the difference between an examinee's actual or obtained score and the theoretical true score counterpart. The *SEM* is a numerical value that is commonly used—and frequently misused—in interpreting and reporting individual test scores and score differences on tests.

This module provides the information needed to adequately interpret test scores using the *SEM*. It should be helpful for persons who must interpret the scores from classroom and standardized tests and who present these interpretations to others. All of the material in this module is based on classical test theory and is appropriate for norm-referenced testing. Criterion-referenced tests and, more particularly, mastery tests are not dealt with specifically in this module.

The major emphasis of this module is on *using* the *SEM*. In order to provide that emphasis, some theoretical aspects are covered but the focus is on such fundamental areas as 1) errors in measurement, 2) what the *SEM* is and is not, and 3) use of the *SEM* and related values in interpreting obtained test scores, score bands, and test score differences.

Leo M. Harvill is the Assistant Dean for Medical Education and a professor at the James H. Quillen College of Medicine at East Tennessee State University, Johnson City, TN 37614. His current research interests are test-taking ability, test anxiety, and clinical performance testing.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education.

Reliability of Measurements

To understand and use the standard error of measurement requires understanding the basic concepts of test reliability, such as true scores and measurement error. These concepts were considered at length in the instructional module by Traub and Rowley (1991).

Obtained, True, and Error Scores

A raw score or *obtained score* on a test is the number of points obtained by an examinee on the test. The main factor influencing test scores is the ability of the examinees in the content area covered by the test. These scores can also be influenced by a number of other factors, e.g., test items which are ambiguous, examinees who are uninterested in doing well on the test, fatigue on the part of the examinees. When these and related factors are present, examiners cannot assume that the obtained scores are an accurate assessment of the examinees' true abilities.

A *true score* represents that part of an examinee's observed score uninfluenced by random events. The term *true score* is a bit misleading because any *systematic* error such as an examinee's reading ability or test-taking skills (test-wiseness) is considered part of the true or unchanging portion of an examinee's observed score. As Julian Stanley (1971) has stated: "As used, true score is not the ultimate fact in the book of the recording angel. Rather, it is the score resulting from systematic factors one chooses to aggregate, including any systematic biasing factors that may produce systematic incorrectness in the scores" (p. 361). Systematic biasing factors such as test-wiseness do not affect test reliability but certainly can negatively affect test validity.

The *error of measurement* or *error score* is the difference between an obtained score and its theoretical true score counterpart. The error score is that part of the obtained score which is unsystematic, random, and due to chance. It is the accumulated effects of all uncontrolled and unspecified influencing factors included in the test score.

Thus, it is possible to express an obtained score by two component parts—a *true* component that represents the examinee's *true* ability and an *error* component that represents chance or random fluctuation,

$$X = T + E \qquad (1)$$

where X = observed score, T = true score, and E = error score. It is possible for the obtained score to be either above or below the true score. As an example, consider guessing by the examinee. If guesses are lucky, the test-taker's obtained score will be above the true score and the error score will be positive. If guesses are unlucky, the test-taker's obtained score will be below the true score and the error score will be negative. In the long run, the positive and negative measurement errors would be expected to cancel each other for any particular examinee since they are random. If they do not cancel out, the factor influencing the scores is probably systematic and not one of chance.

It would be very convenient if the obtained score always equaled the true score for every examinee on every test. If that were the case, there would be no errors of measurement. But such is not the case.

Obtained, True, and Error Variance

Just as the obtained score is equal to the sum of the true score and the error score, the variation in the obtained scores across examinees in the population of interest is equal to the sum of the variation among the true scores and the variation among the error scores. This relationship will hold if the errors are random (as assumed) and do not correlate with the true scores or with each other. This relationship can be expressed as

$$S_X^2 = S_T^2 + S_E^2 \quad (2)$$

where

S_X^2 = variance of observed scores

S_T^2 = variance of true scores

S_E^2 = variance of error scores.

The derivation of Equation (2) can be found in Gulliksen (1950).

Reliability

What does all of this have to do with reliability? Theoretically, the reliability of a test, denoted $r_{xx'}$, is defined as the ratio of the true score variance to the observed score variance:

$$r_{xx'} = S_T^2/S_X^2. \quad (3)$$

Reliability tells us to what extent the observed score variance is due to true score variance. If a test is perfectly reliable, the true score and obtained score variances are equal and the test reliability equals +1.00. The test reliability can also be expressed as

$$r_{xx'} = 1 - S_E^2/S_X^2. \quad (4)$$

Equation (4) shows the relationship between the test reliability and the amount of variance among error scores. It can be seen that as the variance of error scores decreases (in relation to the total test score variance), the test reliability will increase.

Equations (3) and (4) provide technical definitions of reliability, but what does reliability mean in laymen's terms? Synonyms for test reliability are consistency, dependability, and precision. In the *Standards for Educational and Psychological Testing* (1985), reliability is defined as "the degree to which test scores are consistent, dependable, repeatable, that is, the degree to which they are free of errors of measurement" (p. 93). Reliability has to do with score dependability or precision but not with score meaning, accuracy, or validity. Test scores must be dependable or relatively free from random measurement error in order for users to make meaningful inferences about those scores.

Reliability coefficients can be obtained for tests in a number of different ways (see, for example, Traub & Rowley, 1991). These types of reliability coefficients can generally be classified into one of the following categories: stability, equivalence, and internal consistency. In all cases, the reliability of a test is estimated from obtained test scores from a group of examinees. Reliability coefficients range from zero to +1.00.

The reliability of a test is not a fixed value. It will vary among different methods for determining reliability using a single group of examinees and among different groups of examinees using a single method for estimating reliability. The manual for a standardized test may report many reliability coefficients obtained for that test using different methods and different groups of examinees.

While a reliability coefficient for a test can give a good estimate of the extent to which measurement errors may be present or absent in a group, it cannot be used to carry over into individual score interpretation. We cannot use it to determine the effect of measurement error on the obtained test score of an individual examinee. However, the standard error of measurement can be used for this purpose.

Standard Error of Measurement

Defining the Standard Error of Measurement

The standard error of measurement (*SEM*) is defined in the *Standards for Educational and Psychological Testing* (1985) as "the standard deviation of errors of measurement that is associated with the test scores for a specified group of test takers" (p. 94). It is a measure of the variability of the errors of measurement and is directly related to the error score variance (S_E^2) discussed in the previous section. It can be shown with some algebraic manipulation of Equation (4) that

$$S_E^2 = S_X^2(1 - r_{xx'}). \quad (5)$$

The equation for the *SEM* is found by taking the square root of both sides of Equation (5) to get

$$SEM = S_E = \sqrt{S_X^2(1 - r_{xx'})} = S_X\sqrt{(1 - r_{xx'})}. \quad (6)$$

Stated verbally, in order to get an estimate of the *SEM*, subtract the test reliability from one, take the square root of that difference, and multiply the square root value times the standard deviation of the test scores. Some relationships between the *SEM* and the test reliability can be seen from Equation (6). If the reliability of the test is zero, the *SEM* will be equal to the standard deviation of the obtained test scores. If the reliability of the test is +1.00, the highest possible value, the *SEM* is zero. There would be no errors of measurement with a perfectly reliable test; a set of errors all equal to zero has no variability.

The type of reliability coefficient used in calculating the *SEM* can make a difference, both computationally and logically. A long-term stability coefficient (e.g., test-retest with six months intervening) would be expected to be lower than a short-term stability coefficient (e.g., test-retest with two weeks intervening) for the same test. Since a lower reliability estimate will provide a higher *SEM* estimate, the type of reliability coefficient used can have an effect on the magnitude of the *SEM*. Logically, if inferences about individual scores on two different forms of the same test are to be made, it would be appropriate to use an alternate form reliability coefficient in determining the *SEM*. If inferences about individual scores are to be made concerning what the score might be if the examinee were retested with the same test form in six months, it would seem logical to use a test-retest reliability value in calculating the *SEM*. Since each type of reliability coefficient is measuring the effect of different sources of measurement error [character-

istics of the examinees, characteristics of the test(s), administration and scoring of the test(s)], the *SEM* derived from each will have a different meaning.

The reliability coefficient ($r_{xx'}$), the error variance (S_E^2), and the *SEM* (S_E) are all indirect or direct indicators of variation of the errors of measurement. What does the *SEM* provide that the other two do not? The *SEM* allows one to make statements about the precision of test scores of individual examinees; the reliability coefficient does not. The numerical value of the reliability coefficient also depends, to some extent, on the amount of variation among the scores in the group that was tested, but it can be seen in Equation (6) that the *SEM* takes both the group variation (S_X^2) and the reliability coefficient ($r_{xx'}$) into account. As S_X^2 and $r_{xx'}$ tend to increase together, that effect is cancelled out in the calculation of the *SEM* because the increasing S_X is multiplied by the square root of the decreasing quantity ($1 - r_{xx'}$). The unit of measurement for the *SEM* is the same as the unit of measurement for the original test scores; the unit of measurement for the error score variance (S_E^2) is not.

Estimating Obtained Score From True Score Using SEM

If we assume that an examinee's obtained scores on many parallel (interchangeable) forms of a test will vary and will be normally distributed, we would expect the average obtained score to be a good approximation of the examinee's true score and the *SEM* to be the standard deviation of that distribution of obtained scores. These assumptions are appropriate for measurements of this type and were documented centuries ago by Carl Gauss. He was probably the first person to view the bell-shaped or normal curve as a model for depicting random error in measurement. Such a distribution of obtained scores for a single examinee is depicted in Figure 1.

The value at the center of the normal curve (50) represents both the average of an examinee's obtained scores from repeated testing and the examinee's true score. The vertical lines are drawn at intervals representing standard deviation or *SEM* units; in this example, the SEM is equal to two. It can be seen that a majority of the obtained test scores lie between 48 and 52. Using the properties of a normal curve, areas under the curve can be translated into probabilities or likelihoods. It is known that 68 percent of the area of a normal curve lies between one standard deviation below the mean and one standard deviation above the mean. Thus, if an examinee's true score is 50 and the *SEM* is two, the chances are about two out of three (68%) that her obtained score on a single test administration will be between 48 and 52. We could be 95 percent confident that her obtained score would be between 46 and 54, and we could have greater than 99 percent confidence that her obtained score would be between 44 and 56. However, we pay for the added confidence with wider, less precise score bands.

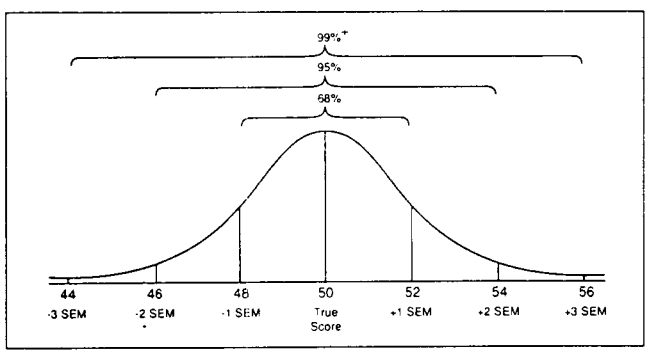

FIGURE 1. *Theoretical distribution of observed test scores obtained from administration of many parallel test forms to single examinee*

There are two problems with this interpretation. The *SEM* value, in practice, is calculated by using values (S_X and r_{xx}) based on all the scores of a group of examinees. But is the single *SEM* (based on *group* data) appropriate for interpreting *individual* scores all along the score continuum? Probably not. (Hold on to that question; it will be answered later in greater detail.) The second problem is that we have gone at the process backwards. We never know the examinee's true score and, if we did, there would be no need to predict her obtained score. We need to be able to predict true scores based upon knowledge of obtained score. Can we use the *SEM* for doing this? The answer is "no" although it is often done that way as an approximation. (This issue is addressed more fully later.) The *SEM* is the standard error of estimate for predicting the obtained score from the true score but it is *not* the standard error of estimate for predicting the true score from the obtained score.

Different SEMs at Different Points on the Score Scale

It has been recognized for some time that the *SEM* derived from Equation (6) is a global or "average" value for the entire test and that the *SEM* is probably different at various points along the score scale for most tests. There has been a growing interest in the *SEM* at specified ability levels or minimum passing levels because of increased use of minimum competency or mastery tests. Test developers and test users are particularly concerned about the *SEM* at a particular cut score used for making decisions as compared to the overall test *SEM*. Feldt, Steffan, and Gupta (1985) provided evidence that the *SEM*, quantified in raw score units, reached a peak in the middle of the score range and that it did vary across score levels. The maximum was often more than twice the minimum. Their findings were consistent across five methods for estimating *SEM*, three tests, and two grade levels. Feldt et al. used methods of estimating *SEM* that were theoretically similar. They were all variations of the same concept: if examinees could be grouped together based on true scores, error variance at each true score would be equal to $\Sigma P_j(1 - P_j)$, where P_j is the difficulty of item j for examinees at that specific true score. The similarity of the empirical findings reflects this underlying conceptual framework. The authors state, "This implies that the standard error of measurement computed by the traditional formula for the test as a whole does not adequately summarize the error propensity of many—perhaps most—examinees" (p. 358).

Estimating SEM *From the Number of Test Items*

In order to calculate the *SEM* using Equation (6), it is necessary to compute the test reliability and the standard deviation of the test scores for a group of examinees. Teachers, test users, or test developers may wish to estimate the *SEM* for a particular test before it is administered or to estimate the number of test items needed to achieve a particular *SEM*. Lord (1959) determined empirically that the *SEM* was directly proportional to the square root of the number of items on the test ($\sqrt{n}$). The correlation between *SEM* and $\sqrt{n}$ was 0.996. These findings were based on 50 objective tests of aptitude and achievement of *moderate difficulty*. The *SEM*s were determined by using the Kuder-Richardson 20, a measure of internal consistency. A good approximation for the *SEM* would be: $SEM \approx 0.432\sqrt{n}$. When the analysis was repeated using the Kuder-Richardson 21 formula for reliabilities, Lord obtained similar results with a correlation of 0.999 and an approxima-

tion for the *SEM* of 0.478 $\sqrt{n}$. The empirical results obtained were reasonable approximations under the conditions presented. To get a rough approximation for the *SEM* of a *moderately difficult* test, given that the reliability used was a measure of internal consistency, multiply the square root of the number of test items by 0.45. If the test is an easy one, with an average score of approximately 90 percent, the multiplier should be closer to 0.3.

Interpretation of Test Scores

Score Bands

A score band is a range of test scores, instead of a single value, and is used in estimating true scores and other test outcomes for test score interpretations. Score bands are sometimes called confidence intervals or confidence bands because they allow us to make probabilistic statements of confidence about an unknown value. Score bands have lower and upper limits on the score scale and provide an estimate that is a range or band of possible test scores. An example of a score band or confidence interval which was used earlier is, "I am 95 percent confident that the examinee's obtained score will be between 46 and 54 (given a true score of 50 and an *SEM* of two)." Score bands can be stated in terms of raw scores (number of items answered correctly), percentage scores, percentile scores, and various standard scores, e.g., stanines, *T* scores, *z* scores, grade equivalent scores.

In this section, only those score bands which are 68 percent confidence intervals will be covered since these are most commonly used in practice. Remember that in a normal distribution:

1. the area (probability) between one standard deviation below and one standard deviation above the mean is 68 percent of the total area under the curve,
2. the area between two (actually 1.96) standard deviations below and two standard deviations above the mean is 95 percent, and
3. the area between 2.58 standard deviations below and 2.58 standard deviations above the mean is 99 percent.

These values can be used to determine score bands for other levels of confidence. The process would be the same as the one illustrated below; only the multiplier (number of standard deviation units) would be changed.

The two score bands to be discussed are the score bands around the following: 1) a true score, to estimate what an obtained score would be, and 2) an obtained score, to estimate what a true score would be.

Score Band Around True Score to Estimate Obtained Score

This score band was presented in the previous section. It would appear that this score band has no practical value or application since the true score is an unknown quantity for any particular examinee. However, this type of score band can be useful for determining appropriate minimum or maximum obtained score cutoffs for making decisions for, for example, program entrance. For example, if it is assumed that a youngster must have a true score IQ of 130 in order to enter a program for the gifted, what minimal obtained score IQ would be acceptable for program entrance? A reiteration of this type of score band is presented here for comparison purposes with the other type of score band.

A 68 percent score band used to estimate an observed test score if the true score were known (or assumed) is given by

$$T \pm (1)(SEM) \qquad (7)$$

where T = examinee's true score, and *SEM* = standard error of measurement. The "±" sign indicates that the *SEM* value should be added to and subtracted from the true score value to obtain the upper and lower limits, respectively, of the score band. If an examinee's true score on a test was 75 and the *SEM* was 7, it could be stated that "the chances are two out of three (68%) that the examinee's obtained score on a single administration of the test would be between 68 (75 − 7) and 82 (75 + 7)." Note that this type of score band is symmetrical around the true score because the same value is added and subtracted to obtain the upper and lower limits.

Directly related to this type of score band is the practical problem of determining how probable a particular obtained score would be if the examinee's true score was some assumed value. For example, what is the probability of a child obtaining an IQ score of 75 if her true IQ score is 70? The answer to this question could be used in setting a maximum cut score or making a decision about retesting examinees in a "gray area" when a decision must be made about entrance into an educational program.

Assume that we have been asked to set a minimum cutoff score for entrance into a program for gifted students. We have been asked to assume that a minimum true score IQ of 130 is required for entrance and that the *SEM* for the test being used is 6 at that score level. If we decide that the minimum score required for entrance into the program is 126, what is the probability that we will *falsely reject* a truly qualified student? Other ways of asking the same question are "What percentage of truly qualified students will be rejected for the program?" and "How likely is it that a person with a true score of 130 will have an obtained score of 126 or less?"

The score of 126 is 4 points below the assumed true score cutoff of 130; it is also 4/6 or .67 of an *SEM* below the value of 130. Using the known properties of the normal curve and a normal curve table, we can determine that 25 percent of the area of the normal curve is to the left of the value of 126 on a normal curve centered at 130 (see Figure 2). Another way of expressing this is to determine that

$$z = (X - T)/SEM = (126 - 130)/6 = -4/6 = -.67$$

The area under the normal curve between the center and .67 *SEM*s below the center is .25 or 25 percent. This can be found from the normal curve table in the back of any statistics text. The area to the left of that same value is also 25 percent since half (50%) of the curve is below the center. Since areas under the curve correspond to probabilities, the probability that an individual with a true score of 130 would be excluded from the program would be 25 percent. Is that probability or percentage too high? If it is, we would have to lower the obtained score cutoff value in order to falsely reject a smaller percentage of truly but minimally qualified candidates. For example, with a cut score of 120, we would falsely reject only 5 percent of the truly but minimally qualified students:

$$z = (120 - 130)/6 = -10/6 = -1.67$$

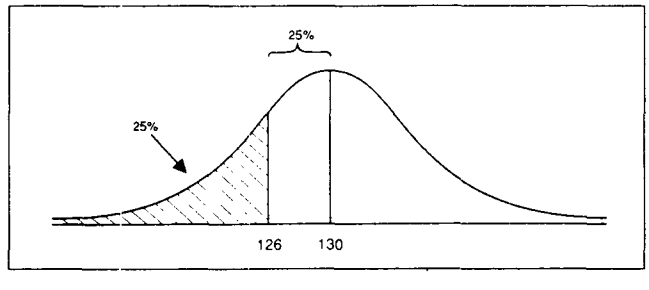

FIGURE 2. *Determining probability for a cut score of 126 with an assured true score of 130*

Five percent of the area of the normal curve is below (to the left of) that point on the normal curve (1.67 *SEM*s below the assumed true score).

Score Band Around Obtained Score to Estimate True Score

Calculating a score band around an obtained test score to estimate an examinee's true score on that test is very popular in educational testing. This approach can also be used for making decisions about placement in exceptional programs. The question might be asked, "If a youngster has an obtained test score of 79 percent, how likely is it that the true score is 80 percent (or higher)?"

This method has been presented by Gulliksen (1950); the equation is

$$[\overline{X} + (r_{xx'})(X - \overline{X})] \pm (1)(S_X)(\sqrt{1 - r_{xx'}})(\sqrt{r_{xx'}}) \quad (8)$$

where

$\overline{X}$ = mean score for an appropriate reference group

$r_{xx'}$ = reliability coefficient

X = obtained test score

S_X = standard deviation of test scores for an appropriate reference group

The value $[\overline{X} + (r_{xx'})(X - \overline{X})]$ in Equation (8) is an estimate of the examinee's true score based upon the obtained score, the mean for the reference group, and the test reliability. If a test has a perfect reliability, ($r_{xx'} = 1$), the estimated true score will be equal to the obtained score.

Here is an example. Bill took a fifth-grade mathematics achievement test and got a score of 79 percent. The test manual indicated that for a group of 1200 fifth-grade boys, the mean percentage score was 73 percent, the standard deviation was 9 percent, and the reliability coefficient (coefficient alpha) was 0.93. Equation (8) can be used to determine a 68 percent score band for estimating Bill's true score:

$$[\overline{X} + (r_{xx'})(X - \overline{X})] \pm (1)(S_X)(\sqrt{1 - r_{xx'}})(\sqrt{r_{xx'}})$$

$$[73 + (.93)(79 - 73)] \pm (1)(9)(\sqrt{1 - .93})(\sqrt{.93})$$

$$[73 + (.93)(6)] \pm (9)(\sqrt{.07})(\sqrt{.93})$$

$$[73 + 5.6] \pm (9)(.26)(.96)$$

$$78.6 \pm 2.3$$

Lower limit of the score band: $78.6 - 2.3 = 76.3$

Upper limit of the score band: $78.6 + 2.3 = 80.9$

There is a 68 percent likelihood that Bill's true score on the mathematics achievement test is between 76.3 and 80.9. The center for this score band is the estimated true score of 78.6, not Bill's obtained score of 79. This is because the test is not perfectly reliable and scores are expected to vary somewhat based on random error. Since the obtained score was above the mean, the true score is assumed to be shifted toward the mean because the test is not completely reliable. Obtained scores below the mean will have predicted true scores shifted toward the mean of the group as well. This is why it is very important in using this equation to use a mean and reliability coefficient that represent a reference group similar to the examinee(s) with which you are dealing. How likely is it that the true score is 80 percent or higher? The value of 80 percent is 1.4 percent above the estimated true score of 78.6 percent. If 1.4 is divided by the *SEM* of 2.3, we find that the value of 80 percent is .61 *SEM* above 78.6 percent. Using a table of normal curve values, we can determine that .2709 of the area of the curve is to the right of .61 *SEM*. This indicates that the likelihood or probability that Bill's true score is 80 percent or above is approximately 27 percent.

The correct procedure described above for calculating score bands around obtained scores has been known for many years and yet it is not mentioned in a number of measurement textbooks. One such text states, "If we add and subtract one standard error of measurement from a given person's test score, we will have the *approximate* range within which his true score actually lies." This is an appropriate statement if we emphasize the word *approximate*. The same author goes on to use an example with an obtained intelligence test score of 110 and an *SEM* of 4.5. He states, "The probability is .68 out of 1.00 that Jane's true score lies between 105.5 and 114.5, i.e., 110 ± 4.5." This author is incorrectly using the first type of score band discussed in this section by replacing the true score (T) in Equation (7) with the obtained score (X) in order to estimate the true score.

Another author uses the two confidence intervals ($T \pm SEM$ and $X \pm SEM$) as if they were completely interchangeable. After calculating an *SEM* value of 3, the author states, "The result indicates that the chances are about two to one that any obtained score on this test will not vary from the true score by more than 3 points. More specifically, the chances are about two to one that the true score of an individual who makes an actual score of 72 on this test will be somewhere between 69 and 75." The first of his two statements is correct; he is correctly using Equation (7). His second statement is not completely accurate; he has replaced T with X in Equation (7).

The *SEM* is an estimate of the variability expected for observed scores when the true score is held constant. To set score bands for true scores when an observed score is held constant, the appropriate standard error is estimated by $(S_X)(\sqrt{1 - r_{xx'}})(\sqrt{r_{xx'}})$ and the interval is centered around the *estimated* true score and not the observed score.

Have these authors who have mixed the score bands in this way committed a serious error? Generally speaking, probably not. Gulliksen was well aware of this procedure and coined the phrase "reasonable limits" for this approximate score band (using $X \pm SEM$ to estimate true score) to indicate that it can be a useful and easy way to estimate true scores *if* precise probability statements are not tied to them (pp. 17–20). The technique of adding the *SEM* to and subtracting it from the obtained score to estimate the true score is a satisfactory estimate *if*, for the reference group of interest, $r_{xx'}$ is reasonably high and if the obtained score of the examinee is not an extreme deviate from the mean of the reference group. If these conditions hold, $X \pm SEM$ provides a good approximation to Equation (8). It is only when either or both of these conditions do not hold that Equation (8) provides an obviously better estimate.

Interpretation of Difference and Change Scores

Is the difference between the verbal ability test scores of Cathy and Don "real" or could it be due to errors of measurement? Is there a difference between John's verbal and mechanical ability test scores? Has there been a "real" change or gain from the preinstruction test score to the postinstruction score for Ginny?

These are the kinds of questions we need to answer when we look at difference, change, or gain scores. It is particularly important to consider reliability and errors of measurement when evaluating the *difference* between two test scores. It is necessary to examine the reliability of differences in making both interindividual and intraindividual test score comparisons.

An important principle to remember is that the difference between two test scores is less reliable than the two individual scores. This is because a difference score contains two sources of measurement error, one from each of the two test scores.

In order to compare the scores of two individuals on the same test, the process of calculating "reasonable limits"

around each score ($X \pm SEM$) and then checking for overlap in the two score bands is often used. If the two score bands do not overlap, the difference is considered to be a meaningful or important one. Another way of stating this same comparison is that a difference equal to or greater than two *SEM*s is a meaningful one. A better statement concerning a two *SEM* difference between scores would be the following: "It is quite likely that the difference between the two scores is a real one and is not likely to have happened by chance alone." In other words, the difference between the two scores is statistically significant. For example, if Cathy's verbal ability score was 52, Don's was 60, and the *SEM* for the test was 3, we could state that the difference between the two scores is significant or real and that Don probably has greater verbal ability than Cathy as measured by this test.

It is important to remember that the *SEM* is not the same at all score levels; the *SEM* for a test score at either extreme in the range of scores could be much lower than the *SEM* for a test score at the middle of the score range. Using the *SEM* calculated for the total test score range may mask or enhance some score differences depending upon where they occur within the score range. In other words, a raw score difference of five points may not be significant in the middle of the score range but it may be at either of the two extremes.

The standard error of measurement for the score difference between individuals *A* and *B* on a single test is:

$$SEM_{A-B} = S_X\sqrt{2 - r_{xx'} - r_{xx'}} = (\sqrt{2})(S_X)(\sqrt{1 - r_{xx'}}) \quad (9)$$

This value is $\sqrt{2}$ or 1.414 times the *SEM* for the test and indicates the contribution of two error sources in the difference scores.

Test publishers often provide test scoring and analysis services. Part of that service includes printing a score profile with confidence bands for individual examinees. These confidence bands are usually derived by using $X \pm SEM$ and then converting the upper and lower raw score limits into percentiles or some standard score. The common interpretation provided by the material accompanying the score profile is that if the bands do not overlap, the difference in the scores is significant. Be wary of statements like this one taken from such a score profile: "When two confidence bands do not overlap, *we can be sure* that the student's performance differed in the two areas." Remember that even if the confidence bands do not overlap, we can only make probabilistic statements about the test scores; we really *cannot be sure* that there was a real difference between the two test scores.

A more precise method to determine if there is a difference between Sue's verbal and mechanical ability test scores would be to use the *SEM* value given in Equation (10) in comparing two test scores for a single examinee. The standard error of measurement for the score differences between any two tests included in the profile for a single individual is

$$SEM_{X-Y} = (S_X)(\sqrt{2 - r_{xx'} - r_{yy'}}) \quad (10)$$

where X and Y are the two tests and $r_{xx'}$ and $r_{yy'}$ are their reliabilities.

The reliability of the difference scores can be determined by

$$r_{DD} = [r_{xx'} + r_{yy'} - 2r_{xy}]/[2(1 - r_{xy})] \quad (11)$$

where $r_{xx'}$ and $r_{yy'}$ are the test reliabilities and r_{xy} is the correlation between the two tests. It is obvious that the difference scores will be very reliable if each of the two tests is highly reliable and there is little relationship between the scores from the two tests. For example, if $r_{xx'} = r_{yy'} = .90$ and $r_{xy} = .20$, then $r_{DD} = .875$. But if $r_{xx'} = r_{yy'} = r_{xy} = .60$, then $r_{DD} = .00$.

It is possible that low reliability may result if the difference score of interest is a change or gain score for a group of individuals, e.g., before and after instruction. Using Equation (11), we can see that if the correlation between preinstruction and postinstruction measures (r_{xy}) is high, the reliability of the gain scores can be quite low. It would seem logical that the preinstruction and postinstruction measures would be similar in content in order to measure the same instructional objectives and, thus, have a high correlation.

Rogosa and Willett (1983) have indicated that perhaps the criticism of gain scores as being *inherently* unreliable has been too harsh. Their arguments show that assumptions made in deriving Equation (11) are probably not correct for gain scores (although they are appropriate for other difference scores). They argue that gain scores *can* be reliable but they do not suggest that gain scores are usually reliable in practice.

What can we conclude about difference scores? Difference scores are less reliable and have larger *SEM*s than single scores. Caution should be used when interpreting difference scores and especially when decisions about individuals follow from those interpretations.

Recommendations for Reporting Test Scores to Examinees

Score bands or confidence bands are the best way to report test scores to examinees or other interested persons. This procedure gives the interpreter a way of expressing the unreliability of the test scores in a nontechnical way. The following points should be considered when constructing, selecting, and interpreting score bands:

1. When calculating score bands, use the appropriate score band for the situation whenever possible. The score band of $X \pm SEM$ provides "reasonable limits" for estimating true score; it provides an adequate approximation when the test reliability is reasonably high and the obtained score for the examinee is not an extreme deviate from the mean of the appropriate reference group. The most appropriate score band for estimating true score from an obtained score is given by Equation (8).
2. When using score bands reported in a test manual, determine how the test publisher calculated the score bands reported in the test norms and what level of confidence was used. It is important to have this information in order to more effectively report the meaning of these scores to others.
3. If you are using "reasonable limits" ($X \pm SEM$) for estimating true scores, be careful with your statements so that they do not imply greater precision than is actually involved. For example, you might say, "It is fairly likely that your daughter's true ability lies between 110 and 120," or "Since these two confidence bands do not overlap, you can be fairly confident that your son's verbal ability is greater than his mathematical ability." Do *not* make such definitive statements as "The chances are *two out of three* that your daughter's true ability lies between 110 and 120," or "Since these two confidence bands do not overlap, *you can be sure* that your son's verbal ability is greater than his mathematical ability." These latter statements are too precise.
4. If the test publisher has provided *SEM* estimates at various test score levels, it would be advisable to base confidence bands on the conditional *SEM*s applicable to the examinees' score levels and not on the *SEM* for the whole test.
5. Use the reliability and *SEM* estimates reported for the reference group which best represents your examinee(s) for calculating confidence bands. Prior to test administration, determine whether an appropriate reference or norm group is mentioned in the test manual.

6. Test users have a responsibility to determine that the information available regarding reliability and measurement errors is relevant for the score interpretations they wish to make. If such relevant information is not available, the test user may need to select another test.

Summation

The standard error of measurement is the standard deviation of errors of measurement that is associated with the test scores for a specific group of test-takers. But, it also takes on different values at varying score levels within a group of examinees. Thus, the *SEM* is a test characteristic estimated from a particular group of examinees, but it is also a test score characteristic varying within the group.

The *SEM* is used to provide score bands or confidence bands around obtained scores to arrive at estimates of true scores which can be used effectively in interpreting test scores to examinees. This type of score band ($X \pm SEM$) should be interpreted appropriately. More precise methods for calculating score bands are available.

Score bands are helpful in interpreting difference scores between test scores taken by a single examinee and between test scores from two or more examinees, but it is important that score bands be used appropriately in these situations.

In selecting a published test, read the test manual to determine if it has reported the reliability, *SEM*, and norms (including confidence bands) for reference groups similar to the examinees you wish to test. Be sure the test manual explains clearly how this information was gathered and how the confidence bands reported in the manual were calculated.

Exercises

Exercise 1

a) Suppose you are told that 92 percent of the obtained score variance for a particular test is estimated to be true score variance based on test scores from 250 ninth-grade students. What is another name or descriptive term for the 92 percent value given?

b) You administered the same test to a group of 150 ninth-graders and found that the variance of the test scores was 81.5. What is your estimate of the error score variance?

Answer to Exercise 1

a) Remember that the ratio of the true score variance to the obtained score variance is the *reliability* of the test; it can be expressed as a proportion (.92) or as a percentage (92%). See Equation (3).

b) Using Equation (4), your best estimate for the reliability of the test is 0.92 (from part a), and the obtained score variance is 81.5. Therefore,

$$r_{xx'} = 1 - S_E^2/S_X^2,$$

$$S_E^2/S_X^2 = 1 - r_{xx'}, \text{ and}$$

$$S_E^2 = S_X^2(1 - r_{xx'}) = (81.5)(1 - .92)$$

$$= (81.5)(.08) = 6.5$$

Exercise 2

a) A mathematics achievement test was administered to 600 high school geometry students. The mean percentage score was 76.3 percent and the standard deviation of the test scores was 12.5 percent. The reliability coefficient (coefficient alpha) was .84. Calculate the standard error of measurement (*SEM*) based on this data.

b) If a student's true score for the mathematics test was 80 percent, what lower and upper percentage scores would contain approximately two-thirds of that student's obtained scores with repeated testing with interchangeable forms of that same test?

Answer to Exercise 2

a) Using Equation (6)

$$SEM = S_X\sqrt{(1 - r_{xx'})} = (12.5)(\sqrt{1 - .84})$$

$$= (12.5)(\sqrt{.16}) = (12.5)(.4) = 5.0 \text{ percent}$$

b) From Figure 1, we can see that about two-thirds (68%) of the area of the normal curve lies between one *SEM* below the true score and one SEM above the true score. One *SEM* below the true score is 80 − 5 = 75. One SEM above the true score is 80 + 5 = 85.

Exercise 3

a) A classroom achievement test consisted of 50 moderately difficult multiple-choice items. Find a rough estimate of the standard error of measurement.

b) A classroom achievement test consisted of 30 easy multiple-choice items. Find a rough estimate of the standard error of measurement.

c) Assume that you wanted to construct a multiple-choice test with moderately difficult items that had an *SEM* of 2. Approximately how many items should you use?

Answer to Exercise 3

a) $SEM \approx 0.45\sqrt{n}$

$\approx 0.45\sqrt{50}$

$\approx (0.45)(7.07) = 3.2$

b) $SEM \approx 0.3\sqrt{n}$

$\approx 0.3\sqrt{30}$

$\approx (0.3)(5.48) = 1.6$

c) $SEM = 2 \approx 0.45\sqrt{n}$; $\sqrt{n} = 2/.45 = 4.44$;

$n = 4.44^2 = 19.75$ or 20 items

Exercise 4

An examinee's obtained score on an intelligence test was 88. For a reference group similar to this examinee, the mean score was 100 and the standard deviation was 15. The test-retest reliability coefficient was 0.72. Calculate the lower and upper limits for the 68 percent score band for estimating the examinee's true IQ score. Use Equation (8). Give a verbal explanation of this score band.

Answer to Exercise 4

Substituting the numerical values into Equation (8), we get

$$[100 + (.72)(88 - 100)] \pm (1)(15)(\sqrt{1 - .72})(\sqrt{.72})$$

$$[100 + (.72)(-12)] \pm (1)(15)(\sqrt{.28})(\sqrt{.72})$$

$$[100 - 8.6] \pm (15)(.53)(.85)$$

$$91.4 \pm 6.7$$

The chances are two out of three that the true IQ score of this examinee lies between 84.7 (91.4 − 6.7) and 98.1 (91.4 + 6.7). How likely is it that this examinee has a true score of 100 or higher? Not very likely. Is it possible that his true score is 100 or higher with an obtained score of 88? It is possible, but not very probable.

Exercise 5

Tom and Mike took an intelligence test which provided verbal, performance, and total scores. The *SEM*s for the verbal,

performance, and total scores are given below. Their scores and the *SEM*s are the following:

	Verbal	Performance	Total
Tom	132	110	125
Mike	115	120	118
SEM	6	7	4

a) Is it likely that the difference between Tom's verbal and performance scores is due to random error? Why or why not?

b) Is it likely that there is a "real" difference between the performance scores of Tom and Mike? Why or why not?

Answer to Exercise 5

a) It is *not* likely that the difference in the scores is due to random error. Placing "reasonable limits" ($X + SEM$) around Tom's verbal score gives a *lower* limit of 126. Doing the same for his performance score provides an *upper* limit of 117. Since there is still a gap of 9 (greater than either *SEM*) between those two limits, it seems likely that the score difference is real.

b) It is *not* likely that there is an actual difference between the two scores. "Reasonable limits" around Tom's performance score are 103 to 117 and comparable values for Mike's score are 113 to 127. Since there is considerable overlap between the two confidence bands, the difference could be due to measurement error.

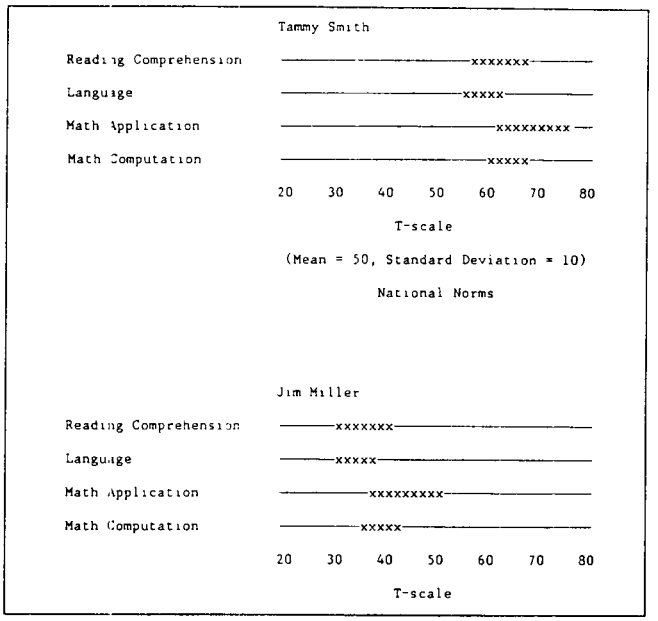

FIGURE 3. *Score profiles for the Omnibus Testing Program*

Self-Test

1. A high school geometry test was administered to 250 students. The mean for the group was 39.6 on the 50 item test; the standard deviation was 4.8. The reliability of the test was estimated to be 0.84 using coefficient alpha. What is the *SEM* based on this data?
2. Assume than an examinee's true score was 80 percent on a test and the *SEM* was estimated to be five percent. What level of confidence (%) would you have that the examinee's obtained score on a single administration of the test would be between 70 percent and 90 percent?
3. A test had a standard deviation (S_x) of 6.4 based on a particular sample of examinees and the *SEM* was calculated to be 3.2. Determine the value of the reliability coefficient used in the calculation.
4. Find an approximate value for the *SEM* for a test consisting of 100 moderately difficult multiple-choice items.
5. Assume that an examinee has a true score of 36 raw score points on a test. The best available estimates of S_X and $r_{xx'}$ for a group of examinees similar to him are 4 and 0.75, respectively. Calculate the lower and upper limits for a 99 percent confidence band to estimate his obtained score.
6. It has been established that a student could be placed in a remedial program if his true score is 80 or less on the entrance test. It is known that the *SEM* is 4 at that score level for the test. If the obtained score cutoff is set at 84, what proportion of students who are truly qualified for the program will be excluded from entering?
7. Susie had a percentile score of 86 on the language subtest and a percentile score of 71 on the mathematics subtest of the School Achievement Test based on national norms. The *SEM* for both subtests was approximately five percentile points. Write an interpretation of this score difference to share with Susie and her parents.
8. An employment test had a mean percentage score of 76 percent, a standard deviation of 12 percent, and a reliability coefficient (alpha) of 0.82 based on the administration of the test to 500 job applicants. If Jim received an obtained score of 64 percent and his 68 percent confidence band had to include the value of 70 percent or above to be hired, did he get the job?
9. Jon got a total score on a college entrance examination of 75; the standard score used has a mean of 50 and a standard deviation of 10 based on a very large sample of college-bound high school seniors. His sister, Angie, scored a 65 on the same test administration. The reliability coefficient for the test is reported to be 0.96. Write an interpretation of this score difference to share with Angie and Jon.
10. Given the score profiles in Figure 3 from the Omnibus Testing Program for Tammy Smith and Jim Miller, answer the following questions: a) How would you describe Tammy's Reading Comprehension and Language scores to her and her parents? b) How would you describe Tammy's Math Application score in comparison with Jim's Math Application score?

Answers to Self-Test

1. $SEM = S_X\sqrt{1 - r_{xx'}} = 4.8\sqrt{1 - .84} = 1.92$
2. Since the true score is 80 and the *SEM* is 5, 70 is two *SEM*s [(2)(5)] below 80, and 90 is two *SEM*s above 80. Approximately 95 percent of the area of the normal curve lies between those limits. Therefore, the level of confidence would be 95 percent.

3. $SEM = S_X\sqrt{1 - r_{xx'}}$

$3.2 = 6.4\sqrt{1 - r_{xx'}}$

$3.2/6.4 = \sqrt{1 - r_{xx'}}$

$.5 = \sqrt{1 - r_{xx'}}$

$.25 = 1 - r_{xx'}$

$r_{xx'} = 1 - .25$

$r_{xx'} = .75$

4. $SEM \approx 0.45\sqrt{n} = 0.45/14\sqrt{100} = (0.45)(10) = 4.5$

5. $T \pm (2.58)(S_X)(\sqrt{1 - r_{xx'}})$

$36 \pm (2.58)(4)(\sqrt{1 - .75})$

$36 \pm (2.58)(4)(\sqrt{.25})$

$36 \pm (2.58)(4)(.5)$

36 ± 5.2

Upper limit = 36 + 5.2 = 41.2
Lower limit = 36 − 5.2 = 30.8

6. We know that 68 percent of the area of the normal curve lies between one *SEM* below the true score and one *SEM* above the true score. Since the curve is symmetrical, half of that area (34%) lies between the true score and one *SEM* above the true score. Since we also know that one-half (50% of the area under the curve) is above the true score, by subtraction we find that 16 percent of the area of the curve is above or to the right of one *SEM* above the true score. Therefore, we can state that 16 percent of truly qualified students (with true scores of 80 or less) will be excluded if a cut score of 84 is used.

7. Susie performed well above the national average on both subtests. Her mathematics score is as good as or better than 71 percent of students at her grade level across the country while her language score was better than 86 percent of those same students. If she were to take tests similar to these many times, we would expect her mathematics percentile score to be between 66 and 76 and her language percentile score to be between 81 and 91. Since those two confidence bands do not overlap, it is likely that the difference between the two subtest scores is a real difference and not due to chance factors. Her language skills appear to be stronger than her mathematics skills.

8. $[\bar{X} + r_{xx'}(X - \bar{X})] \pm (S_X)(\sqrt{1 - r_{xx'}})(\sqrt{r_{xx'}})$

$[76 \pm (.82)(64 - 76)] \pm (12)\sqrt{1 - .82})(\sqrt{.82})$

$[76 \pm (.82)(-12)] \pm (12)(\sqrt{.18})(\sqrt{.82})$

$[76 - 9.84] \pm (12)(.42)(.91)$

66.2 ± 4.6

Upper limit = 66.2 + 4.6 = 70.8
Lower limit = 66.2 − 4.6 = 61.6
The confidence band includes 70 percent so Jim got the job.

9. $SEM = S_X\sqrt{1 - r_{xx'}} = 10\sqrt{1 - .96} = (10)(\sqrt{.04} = (10)(.2) = 2.0$
Jon, your score on the test is unusually high compared with college-bound seniors. We would expect your score to be between 71 and 79 (approximate 95% score band) with many repeated testings. Angie, your score is good compared with other college-bound seniors. We could feel very confident that your score would be between 61 and 69 with repeated testing. It does appear that the difference in your two scores is a real difference because those confidence bands do not overlap. The difference is probably not due to chance.

10. a) Tammy's Reading Comprehension and Language scores are both above average when compared with students across the country at her grade level. The two score bands show considerable overlap; this indicates that the two scores are approximately equivalent and shows approximately the same level of functioning in the two areas.

b) Tammy's Math Application score band does not overlap with that of Jim. This would indicate that there is probably a real difference in their test scores and that Tammy has greater ability in applying mathematical concepts than Jim.

References

AERA, APA, & NCME (1985). *Standards for educational and psychological testing.* Washington, D. C.: American Psychological Association.

Feldt, L. S., Steffan, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9,* 351–361.

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

Lord, F. M. (1959). Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement, 19,* 233–239.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20,* 335–343.

Stanley, J. C. (1971). Reliability. In R. Thorndike, (Ed.), *Educational Measurement,* (2nd ed., pp. 356–442). Washington, D. C.: American Council on Education.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational measurement: Issues and practice, 10*(1), 37–45.

Annotated References

AERA, APA, & NCME (1985). *Standards for educational and psychological testing.* Washington, D. C.: American Psychological Association.

This publication provides criteria for the evaluation of tests, testing practices, and the effects of test use. Standards pertaining to reliability and error measurement are found on pages 19–23.

Anastasi, A. (1988). *Psychological testing (6th ed.).* New York: Macmillan Publishing Company, Inc.

This well respected text presents an excellent discussion of the standard error of measurement including such topics as score bands, reasonable limits, variation of *SEM* with different ability levels, and interpretation of score differences on pages 133–137.

Brown, F. G. (1983). *Principles of educational and psychological testing (3rd ed.).* New York: Holt, Rinehart, and Winston.

This beginning text provides a good, basic introduction to the topics of reliability and measurement error with a brief discussion of score bands.

Mehrens, W. A., & Lehmann, I. J. (1984). *Measurement and evaluation in education and psychology (3rd ed.).* New York: Holt, Rinehart, and Winston.

This is an excellent textbook in the area of educational and psychological measurement which presents a very good introduction to both reliability and error of measurement on pages 266–287.