

Clarification on Some Misconceptions about Conditional Standard Errors of Measurement

Gregory M. Hurtz
California State University, Sacramento

In Hurtz, G. M. (Chair), *Integrating conditional standard errors of measurement into personnel selection practices*. Symposium presented at the Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA, April 2008.

Abstract

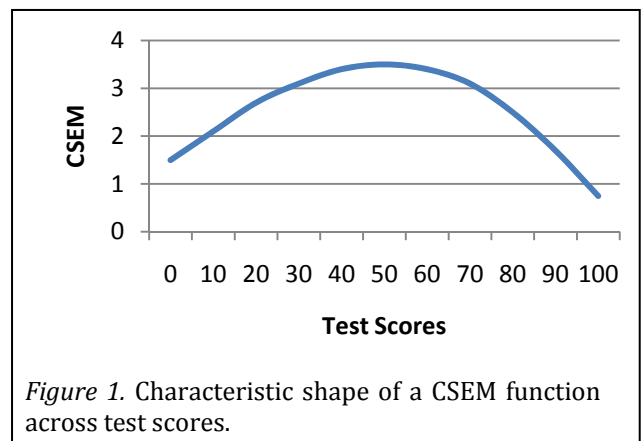
As conditional standard errors of measurement (CSEMs) have gained increasing attention in personnel selection, four misconceptions have arisen that deserve clarification. First, it is often stated that tests are “more reliable” near the endpoints of their score distributions. Second, “true scores” are often discussed as if they are direct indexes of the latent trait being measured by the test. Third, it is frequently stated that classical testing theory only provides static SEM formulas that do not change across test scores. Fourth, it is often observed that classical test theory CSEMs contradict item response theory CSEMs in their shape. Clarification is provided on each of these issues in hopes of helping to advance our understanding of CSEMs as they are increasingly incorporated into personnel selection practices.

Conditional standard errors of measurement (CSEMs) are gaining increasing attention in personnel selection (Biddle, 2006; Bobko & Roth, 2004; Bobko, Roth, & Nicewander, 2005; Harvey, Aguinis, & Gibson, 2007; Harvey, Aguinis, & Wagner, 2007). Recent testing standards (APA/AERA/NCME, 1999) have begun to require or at least strongly recommend their use, placing them in the required toolkit for many personnel selection researchers and practitioners. As the personnel selection field has begun to discuss and study this topic, some misconceptions have arisen that deserve clarification at this early stage of development and implementation of CSEM methodology for the selection context. The goal of this paper is to raise awareness and stimulate discussion of these potential misconceptions and ultimately to help “tighten up” the language we use in our exploration of CSEMs as they gain increasing importance in our discipline.

Misconception 1: Tests are “more reliable” at their upper and lower extremes.

The degree to which one agrees that this is a misconception will undoubtedly depend on one’s focus. If test scores are taken at face value one is likely to be alarmed by the suggestion that this statement could be construed as inaccurate. This is because the classic shape of the CSEM function for observed scores found in virtually all CSEM methods is an inverted-U shape (e.g., Feldt, Steffen, & Gupta, 1985; Qualls-Payne, 1992) as depicted in Figure 1. This function reveals that errors in measurement are larger near the central tendency of the score distribution and smaller near

the endpoints. It seems logical to interpret this observation as indicating that scores are more reliable toward the endpoints. This, however, is impossible according to the linear model that underlies classical testing theory (CTT), and as will be demonstrated later is contradictory to what measurement theory predicts with respect to the shape of the error function.



To understand the problem noted in this misconception, the CTT paradigm needs to be framed in terms of its linear regression roots. CTT is based on the theoretical linear regression of true scores on observed scores; or stated in another way, observed test scores are seen as predictors of test-takers’ true scores. The reliability coefficient is a function of the slope of the line in this regression. The square root of the reliability coefficient (a.k.a., the “reliability index”) is the estimated correlation between observed scores

and true scores (Nunnally & Bernstein, 1994, p. 221), and is therefore equal to the standardized slope (beta weight) of this prediction line. There is only one slope for a line, and hence there can only be one reliability coefficient under this model. To say scores are “more reliable” near the endpoints suggests there are multiple reliabilities for a test which cannot be the case under CTT.

Then what can we make of the lowered CSEMs near the endpoints in this linear regression model? CSEMs are errors, or residuals, in predicting true scores from observed scores and using the linear regression terminology the presence of varying CSEMs across scores is an indicator of heteroscedasticity in the errors. CSEMs, then indicate a potential weakness in the linear model (a violation of one of its main assumptions) but it is a linear model nevertheless with a single slope/reliability.

To further demonstrate, if we consider the classic formula for an SEM:

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

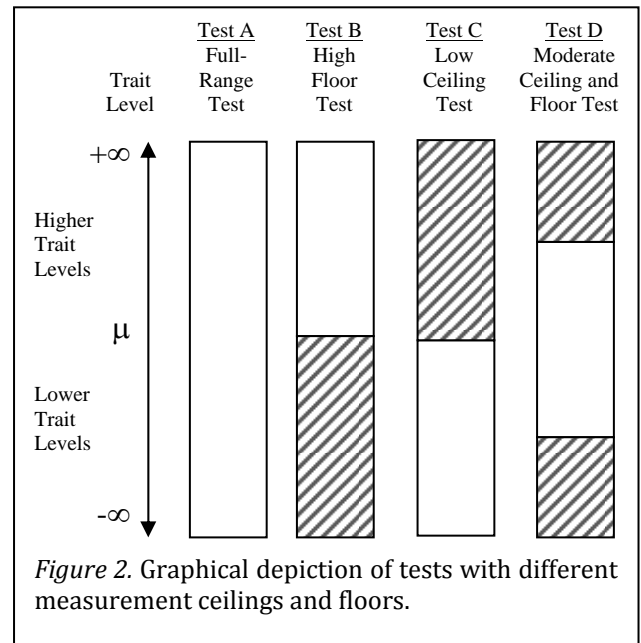
we see that it is a function of two things: The variance (or rather the standard deviation, σ_x) of scores and the reliability (r_{xx}) of the test. Since under CTT the reliability must be constant for a given test for reasons described earlier, the change in SEM across test scores must be a function of different variances across test scores – or, again, of heteroscedasticity. It is potentially misleading to say the test is more reliable near the endpoints. The only allowable interpretation under CTT is that there is less variance in observed scores near the endpoints. Consistent with the current position, it was this reduction in variance that Bobko and Roth (2004; Bobko et al., 2005) focused on in their explanation of the typically smaller CSEMs near the test endpoints.

This reduction in variance is a function of the ceiling and floor on the measurement scale, and can therefore be viewed as a measurement artifact or a limitation of the test. This suggests that the presence of lowered CSEMs calculated near the endpoints may be a sign of a flaw in the test for the particular purpose it is being used. Variance is in fact a good thing in psychometric theory; without variance you cannot differentiate people. The reduction in variance near the endpoints of the measurement scale is a manifestation of the limitations of the scale for providing further differentiation among people scoring near the ceiling or floor.

To illustrate further, consider the depiction in Figure 2. Along the left side is the theoretical “latent trait” continuum of the construct being measured. Four different tests are then depicted that differ in

their ceilings and floors. Test A is a theoretical test that contains items with difficulty levels throughout the range of the latent trait so no test taker ever reaches the ceiling or floor. Test B on the other hand was developed to maximally discriminate among above-average test takers and has a high floor (the cross-hatched area). Test C targets below-average test takers and has a low ceiling, while Test D has a moderately low ceiling and a moderately high floor so that it is targeting the middle range of the trait.

To illustrate why reduced variance at the endpoints of the test is a potential weakness and not a strength, consider the scores that test-takers would receive on these different tests. All test-takers who are above average ($>\mu$) in the latent trait would tend to receive the same score (the maximum score) on Test C, so that the empirically-estimated CSEM on this test in that score region would be very low. Over repeated test administrations this score would in fact be very consistent (and possibly labeled “more reliable”) for these individuals. These same individuals, however, would achieve different scores relative to one another on Tests B and D. Test D would have reduced variance in comparison to Test B which would fully differentiate among people in the above-average range.



If lower CSEMs are taken to suggest the test is more reliable, then Test C would be the most reliable for above-average test takers even though it is completely incapable of detecting differences among these individuals! Likewise, Test D would be more reliable than Test B even though its ability to detect

these differences is limited. Clearly Test B is the best for the above-average group but would be ranked worst in terms of its reliability if reliability is defined according to the size of the estimated CSEM.

Practically speaking, the problems illustrated in Figure 2 are only problems for certain selection strategies. Under top-down selection there is clearly a problem with Test C, Test D is limited, and Test B is probably the most practical (Test A unnecessarily differentiates among people far below average who would probably never be hired). For tests with a minimum cutoff score near the mean, Test D might be the most practical and the low ceiling/high floor will not be a problem. There is no need to differentiate among trait levels that are outside the range where hiring decisions will be made. The main point here is that it is not necessarily a good thing if the relatively smaller estimated CSEMs for a test occur where decisions will be made. This might indicate the test is not appropriately calibrated to the applicant population.

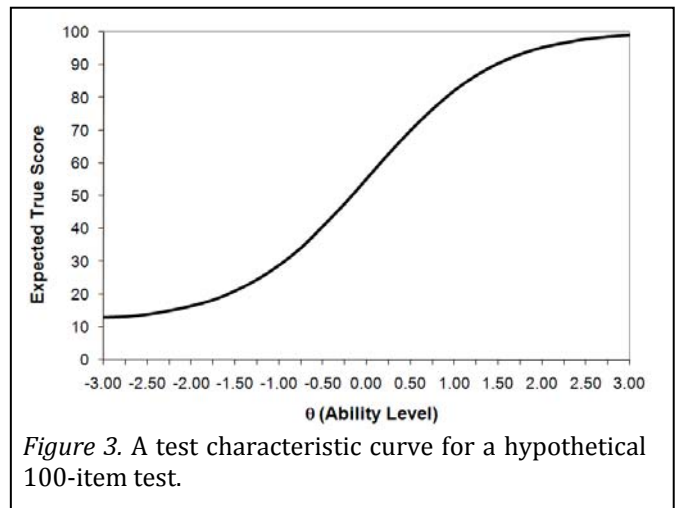
It should be noted that a recent article by Raju, Price, Oshima, and Nering (2007) presented a “conditional reliability” index that is based off estimates of CSEMs for specific test scores. Their index is basically a method for standardizing the CSEMs so that they may be compared across tests with different scoring metrics. They state that their index “does not conform to the usual definition of score reliability” (p. 173). While their explanation of this statement is based on different reasoning than the position stated here, the arguments here likewise suggest that Raju et al.’s conditional reliability index does not conform to the usual definition of reliability, which under CTT must be non-conditional or constant as explained earlier. Raju et al.’s index appears to be potentially very useful in interpreting CSEMs but perhaps it should simply be called a standardized CSEM – a phrase that Raju et al. use synonymously with “conditional reliability” – in order to avoid confusion.

Misconception 2: True scores are direct indexes of latent trait levels.

A “true score” is typically defined as the average score for a person if he or she were to take the test repeatedly (Nunnally & Bernstein, 1994). Allen and Yen (1979), for example, define true scores as “the mean of the theoretical distribution of X scores that would be found in repeated independent testing of the same person with the same test” (p. 57). From these definitions it is apparent that true scores (τ) are estimates of scores *on a particular test*, and are only indirect measures of the underlying latent construct (θ) the test is designed to measure. True scores and

observed scores are both tied to the operational definition of the construct; the weaknesses of that operational definition (e.g., a low ceiling or high floor as depicted in Figure 2) are inherited by the true score metric.

This fact is not shared by other measurement theories such as item response theory (IRT) and the broader class of confirmatory factor analysis models. In these more modern measurement theories there is a clearer distinction between the construct and the operational definitions (i.e., the items or indicators) of that construct. The distinction here can be made very clear by viewing one of the many useful features of IRT – the test characteristic curve. Figure 3 shows a curve for a hypothetical 100-item dichotomously scored test, with a minimum possible score of 0 and a maximum possible score of 100. The horizontal axis displays IRT-estimated latent trait (ability) levels which are typically scaled in the standard score metric and denoted θ . This estimate is asymptotic, and although the figure only displays values from -3 to +3 the scores can theoretically range from negative to positive infinity.



The vertical axis of this graph shows how the latent trait scores translate to the true score scale. The vertical axis is the expected true score (τ) *on the test* that corresponds to each level of the latent trait. There is a clear distinction here between the trait levels and the corresponding true scores. The true scores inherit the ceiling and floor of the operational definition; they have a maximum and minimum value.

This is relevant to the CSEM topic because of the shape of the curve near the ceiling and the floor. Just as Figure 2 showed that individuals beyond the measurement ceiling or floor of a test would tend to achieve the same score on the test even if their

underlying trait levels differed, Figure 3 likewise shows that individuals in the upper and lower regions of the trait tend to have relatively similar expected true scores on the test *even if their trait levels differ widely*. For example, the curve starts to level off rapidly in the positive direction beyond a trait score of +1.00 and flattens fairly quickly. When trait levels are scaled on the true score scale there is little differentiation (little variance, resulting in a low CSEM) among test-takers who have markedly different trait levels. This hypothetical test would be incapable of distinguishing among the “best of the best” in an applicant pool and under top-down selection it would lack the ability to differentiate between the people that decision-makers are looking at most closely for making job offers.

If we use the non-linear model in Figure 3 to revisit Misconception 1, we can see again why we should avoid stating that tests are more reliable near their endpoints. If reliability is the slope of the function that relates observed scores (which in the IRT context are latent ability estimates) to classical true scores, then Figure 3 shows that even when we move beyond the simplistic linear CTT model this conditional slope is flatter, not steeper, at the extremes of the latent trait. Flatter slopes of course would mean lower correlations in those score regions between changes in latent ability and changes in true scores – or in other words lower (not higher) “conditional reliabilities” in those regions. This perspective critiques the issue from the standpoint of the slope of the function; the shape of the conditional errors or residuals about that function provide another angle, as will be discussed in the next section.

Misconception 3: Classical Testing Theory only Provides a Static SEM Formula.

While Figure 1 clearly shows the typical inverted-U shape of the empirically-estimated CSEM function and Figures 2 and 3 demonstrate why CSEMs will tend to be smaller near the measurement ceiling and floor of the test, psychometric theory and practice based in CTT has tended to involve only static SEM formulas that do not allow for different error bands across different test scores. This has led to a common critique of CTT that it does not allow for different SEMs. Embretson and Riese (2000), for example, in their list of 10 old and new “rules” of measurement list this as Old Rule 1: “The standard error of measurement applies to all scores in a particular population” (p. 15). This statement does in fact accurately reflect the history and common practice of CTT, but once again if we frame CTT through the lens of its linear regression roots we can see that CSEM methods can and do exist

under the linear model that serves as the basis for CTT.

Recall that the linear regression perspective on measurement posits that true scores are predicted from observed scores. The standard error of the estimate from this regression is the standard error of measurement for true scores:

$$2) \sigma_{\tau,x} = \sigma_{\tau} \sqrt{(1 - r_{xx})}$$

where $\sigma_{\tau} = \sqrt{(\sigma_x^2)(r_{xx})}$. Note first the similarity between Equation 2 and Equation 1. The typical SEM for observed scores (Equation 1) has the same basic structure as this regression-based formula for the SEM for true scores (usually appropriately called the standard error of estimate, or SEE, in this context). Since the reliability coefficient (r_{xx}) is the square of the reliability index (i.e., the correlation between true and observed scores) the missing power of 2 on the r_{xx} in Equation 2 is simply a matter of notation.

Since $\sigma_{\tau,x}$ is a standard error of estimate in the same sense as regression analysis which has a typical parallel formula of the form:

$$3) s_{y,x} = \sigma_y \sqrt{(1 - r_{xy}^2)}$$

then it stands that we can adjust $\sigma_{\tau,x}$ in the same way that $s_{y,x}$ is adjusted in the context of establishing prediction intervals in linear regression. Conditional standard error of estimate formulas in the regression literature (e.g., Pedhazur, 1997) are used to establish confidence intervals around predicted scores (Y) from a regression model that differ across levels of X:

$$4) s_{\mu'} = s_{y,x} \sqrt{\left[\frac{1}{N} + \frac{(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]}$$

$$5) s_{y'} = s_{y,x} \sqrt{\left[1 + \frac{1}{N} + \frac{(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]}$$

Both of these equations take the “static” standard error of estimate ($s_{y,x}$) and apply a nonlinear correction that has the effect of adjusting the width of the prediction interval in accordance with the deviation of an individual score on X from the mean of X. Equation 4 is used when predictions are to be made with respect to the *mean* of predicted scores, while Equation 5 is used when predictions are to be made for *individual* scores of specific persons.

From these formulas, we can extrapolate two conditional SEE (CSEE) formulas under CTT:

$$6) CSEE_{\mu'} = \sigma_{\tau.x} \sqrt{\frac{1}{N} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

$$7) CSEE_{\tau} = \sigma_{\tau.x} \sqrt{\left[1 + \frac{1}{N} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right]}$$

At least three observations are noteworthy from these formulas. First, we can make separate estimations of CSEEs for the mean true score of all applicants at each observed score (Equation 6) versus CSEEs for each individual applicant who achieves a given observed score (Equation 7). Equation 6 might be useful in selection research evaluating the big picture of what tends to happen over the long run with respect to selection decisions from a given test, while Equation 7 is more appropriate when the focus is on individual selection decisions. The CSEE resulting from Equation 6 will usually be substantially smaller than that resulting from Equation 7.

Second, there are two components to the adjustment to the SEE in both Equations: The distance of the score from the mean, and the sample size. In Equation 6, when a score is at the mean the adjustment is a function of sample size only. The same is true of Equation 7 except that the adjustment is larger. In either equation, the larger the deviation of X from the mean, the greater will be the adjustment.

Third, however, the adjustment factor resulting from the deviation of X from the mean results in *larger* CSEEs toward the extremes of test scores rather than less error toward the endpoints which is the typical CSEM finding as depicted earlier in Figure 1. This pattern can be viewed in Figure 4 which uses Equation 7 to compute CSEEs for percent-correct scores on a 55-item intermediate statistics examination with a mean score of approximately 77.59%, an SD of 12.44, and r_{xx} of .83 based on an N of 43 test-takers. This shape runs counter to all current methods of estimating CSEMs empirically in CTT which result in the classic inverted-U shape.

The fact that the linear regression model that underlies CTT predicts *more* error with greater deviations from the mean while empirically-estimated CSEMs find *lower* errors toward the extremes appears to be the effect of the violation of the homoscedasticity assumption of this linear model that results from the compression of observed variance near the measurement ceiling and floor. These theoretical predictions state that conditional measurement error of true scores (CSEE) will be U-shaped; the empirical findings such as those presented by Feldt et al. (1985) and Qualls-Payne (1992) state that conditional measurement error of observed scores (CSEM) is inverted-U shaped. This contradiction in shape is

addressed further in the context of the next misconception.

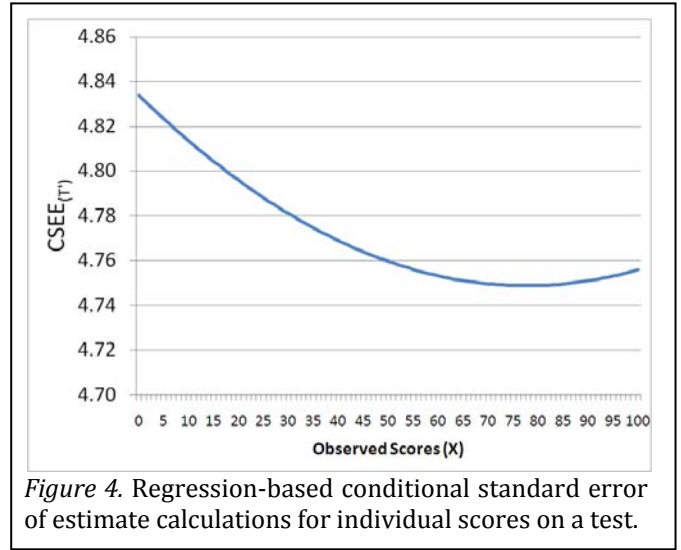


Figure 4. Regression-based conditional standard error of estimate calculations for individual scores on a test.

Misconception 4: Item response theory CSEMs contradict classical test theory CSEMs in their shape.

Like the theoretically predicted shape of true score measurement error, IRT ability CSEMs are typically U-shaped rather than inverted-U shaped (although the specific pattern really depends on the shape of the test information curve). In IRT the amount of error at a given ability level has an inverse relationship with the amount of test information at that ability level. Most often information is highest somewhere in the mid-range of the ability scale and tapers off toward the extremes. The CSEM on this scale is simply the reciprocal of the square root of the information, so that the peak of information is the low point of the CSEM. Figure 5 shows this relationship in a hypothetical test. Where the information function (Panel A) is highest, the standard error function (Panel B) is lowest and vice versa. Ability estimates toward the tails of the ability distribution tend to have the highest, not the lowest, errors.

Here once again we have a model-based error function that is opposite the typical empirically-estimated CSEM under CTT. Raju et al. (2007) recently demonstrated this opposite pattern comparing CTT to IRT methods, noting opposite patterns in both the CSEMs and the “conditional reliability” values derived from those CSEMs. Likewise comparing IRT to CTT methods Harvey, Aguinis, and Wilson (2007) noted the “...paradoxical (from an IRT perspective) conclusion that the test produces its *lowest* precision where the *highest* test information occurs” (p. 4). They later

demonstrated this contradictory shape with both real and simulated data and noted that “Obviously, it cannot be the case that a test provides both its highest and lowest levels of measurement precision in the same range of scores...” (p. 8). Through their Monte Carlo simulation Harvey, Aguinis, and Wilson (2007) demonstrated that true measurement error follows the U shape, and not the inverted-U shape that typical CSEM methods produce.

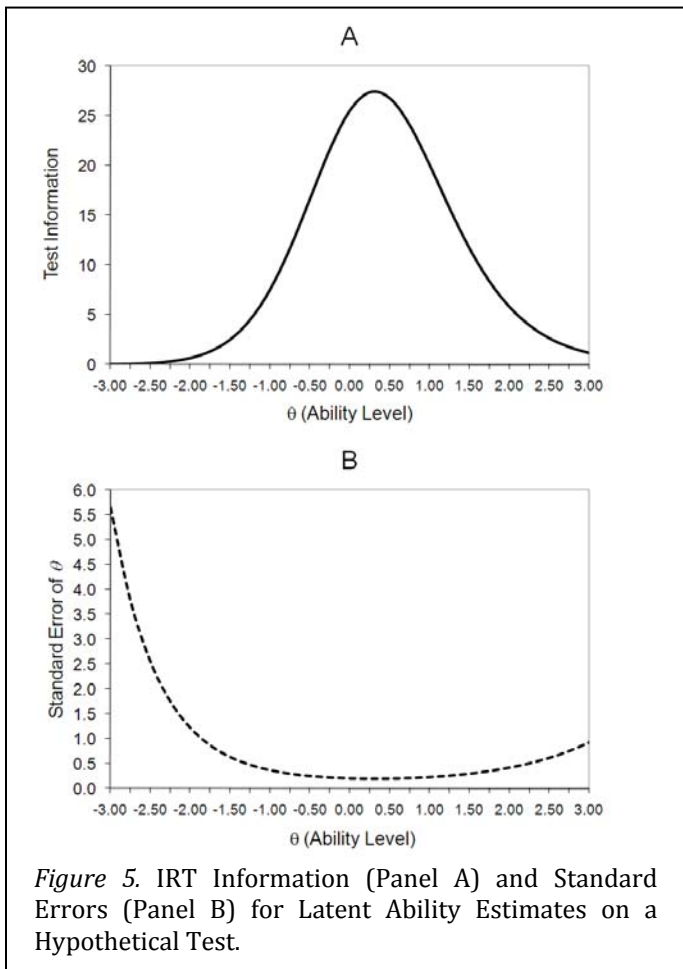


Figure 5. IRT Information (Panel A) and Standard Errors (Panel B) for Latent Ability Estimates on a Hypothetical Test.

As demonstrated previously, the linear model underlying CTT actually does predict less measurement error for true scores in the middle of the distribution and more error as scores deviate from the middle. Therefore, if we are talking about “true measurement error” then both measurement models, at least in theory, make this same prediction. This is one reason I call the apparent contradiction between CTT and IRT a misconception.

However, the fact still remains that the common methods of estimating CSEMs from empirical examinee data produce the opposite pattern. Once again, this pattern appears to be an artifact of the measurement ceiling and floor which restrict observed

variance in their vicinity. As clarified earlier, “true scores” in CTT ultimately inherit the ceiling and floor of the measurement instrument, and it is these limits on the range of scores that produce a compression of errors at the endpoints despite the linear regression model’s theoretical predictions of U-shaped error.

Figure 3 displayed the relationship between the ability (θ) scale and the true score (τ) scale represented as the test characteristic curve in IRT. Figure 6 reproduces that curve but demonstrates the compression of high latent ability estimates when translated to the expected true score scale since individuals with these high latent ability scores reach the measurement ceiling of the operational definition. The Figure reveals that anyone with an ability level falling near the upper extreme (ranging from +2 to positive infinity in Figure 6) will be assigned a relatively restricted range of values on the true score scale. That is, they will max out on the test score range. The variance in scores on the metric that observed and true scores follow will be reduced among the extremely high ability test takers in comparison to those in the middle region because there is simply less room to vary along the test score continuum, even though their true trait levels may vary widely.

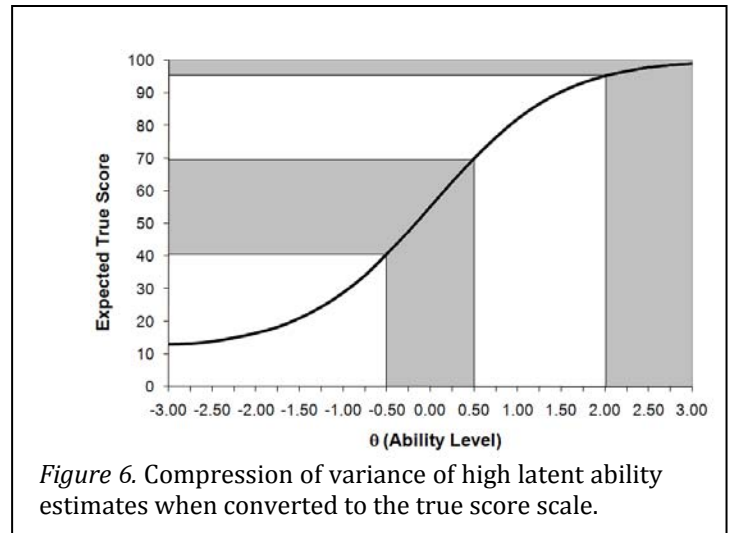


Figure 6. Compression of variance of high latent ability estimates when converted to the true score scale.

This reintroduces the apparent paradox where large errors in IRT will translate to small errors in CTT, but it provides a logical explanation founded in the limitations of the operational definition of the trait and the use of the less sophisticated measurement model provided by CTT. It also demonstrates that the paradox is a function of which metric one is talking about. CSEMs on the θ metric cannot be directly compared to CSEMs on the τ metric because they are different scales of measurement. Harvey, Aguinis, and Wilson (2007) as well as Harvey, Aguinis, and Gibson

(2007) appear to have made this type of improper comparison when noting the apparent paradox, as they compared IRT information-based CSEMs (on the latent ability scale) to the CSEMs produced for the same test using CTT methods and a different IRT method linked to the true score scale. In order to properly compare IRT information-based CSEMs to those based in CTT, one must first do the translation of latent ability estimates to the expected true score scale. Only after placing them on a common scale can their shapes and sizes be properly compared.

Resolution through Terminology

Much of what is discussed in this paper may be viewed at a surface level as matters of semantics. A pure pragmatist might say “if test scores are more consistent and stable for top-scorers in comparison to those near the mean, then how can you suggest that the test is NOT more reliable for top-scorers?” It is in fact hard to argue with this logic at the level of the test (as opposed to latent trait estimates) and with common-language usage of the word “reliability.”

However, as a science we need our technical terminology such as that involved in psychological measurement to have consistent meanings and interpretations. When we speak of the reliability of a test, this term must be universally interpreted in the same manner and we cannot have apparent paradoxes where a test is both the most and least reliable in the same location on the measurement scale. We cannot draw conclusions about test reliability that need to be qualified with descriptions of what is meant by “reliability” in the context of a particular situation.

This issue and most of the misconceptions noted in this paper could be eliminated through more careful use of our technical language. I propose the following recommendations to clean up the language of CSEMs:

- (1) The term “reliability” should only be used to describe the test as a whole, across the entire range of scores. The term “conditional reliability” or suggestions of multiple reliabilities at different test scores should be avoided. To support this suggestion, reliability should be thought of in terms of an R^2 value from a regression analysis. It is an index of the ability of the entire function throughout the range of scores (whether it is a linear function as in CTT or a nonlinear function as in IRT) to minimize the errors in estimating true scores *or* latent trait levels from observed examinee data.
- (2) The term “precision” should be used when discussing CSEMs, as the focus is on the precision

and accuracy of estimates of particular scores. Precision can, and typically does, vary across test scores.

- (3) Discussion of the precision of true score estimates (τ) should be distinguished from discussion of the precision of latent trait level estimates (θ). As demonstrated in Figures 2 and 6 as well as elsewhere as cited in this paper, estimates of τ may be quite precise near the test’s ceiling while estimates of θ may be quite imprecise at the same location. If one is clear about the metric and what is being estimated, then the noted paradox disappears.

Practical Implications

While much of this paper may seem overly academic and esoteric to personnel selection practitioners, there are several take-home messages that are of potential practical value.

- (1) When discussing CSEMs one needs to be clear about which metric they are talking about. Are they CSEMs on the observed score (CTT) scale or the latent ability (IRT) scale? Attention to this question can avoid confusion and seemingly contradictory findings.
- (2) The typical finding of smaller CSEMs for higher-scoring test-takers relative to those near the mean should *not* be said to indicate that the test is more reliable at assessing their levels of the trait being measured. Perhaps their observed test scores can be expected to be highly consistent over repeated test administrations but this is likely due to their reaching the ceiling on the test. The reduced variance actually leads to a reduction in the ability to precisely distinguish among true levels of the underlying trait in this group of test-takers.
- (3) If considerably smaller CSEMs are found near the test ceiling and top-down selection will be used, this should raise suspicion of compressed scores and the test should be revised to raise the ceiling (i.e., make the test more difficult). This will allow the organization to better differentiate among the top candidates for the job.
- (4) If test score banding will be used, the width of the CSEMs which are used to define the width of the score bands can potentially influence which test-takers will be considered equivalent for selection purposes. Appropriate calibration of the test for the applicant population, such as setting a

sufficiently high ceiling, can therefore affect the width of the bands. Spuriously low CSEMs that are the result of a ceiling effect will yield spuriously narrow bands, while a test with a higher ceiling and larger CSEMs would have relatively wider bands.

- (5) If a cutoff score is used on the test that falls somewhere in the middle of the score distribution, then there is probably little need to worry about the ceiling and floor of the test. The test should be designed to target the level of the trait where decisions will be made, and if fine distinctions among the top-ranked test-takers will not influence final hiring decisions then it is largely unnecessary to make those distinctions.

References

- AERA/APA/NCME (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Allen, M. & Yen, W. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- Biddle, D. (2006). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing* (2nd ed.). Burlington, England: Gower.
- Bobko, P. & Roth, P. L. (2004). Personnel Selection with top-score-referenced banding: On the inappropriateness of current procedures. *International Journal of Selection and Assessment*, 12, 291-298.
- Bobko, P. Roth, P. L., & Nicewander, A. (2005). Banding selection scores in human resource management decisions: Current inaccuracies and the effect of conditional standard errors. *Organizational Research Methods*, 8, 259-273.
- Embretson, S. E., & Riese, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Feldt, L., Steffen, 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.
- Harvey, R. J., Aguinis, H., & Gibson, S. G. (2007, April). *Impact of IRT-based top-score banding on ASVAB minority selection ratios*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New York.
- Harvey, R. J., Aguinis, H., & Wagner, T. (2007, April). *Using IRT to produce more accurate and wider test-score bands*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New York.
- Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed). New York: McGraw Hill.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed). Wadsworth.
- Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement*, 29, 213-225.
- Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*, 31, 169-180.