A Comparison of Two Methods for Computing IRT Scores From the Number-Correct Score

Applied Psychological Measurement

2015, Vol. 39(8) 643–655

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621615601081

apm.sagepub.com

W. Alan Nicewander¹ and E. Matthew Schulz¹

Abstract

Two estimates for item response theory latent trait scores (θ) based on the summed, number-correct score, X , were compared: (a) the so-called test characteristic curve (TCC) estimates, θ_{TCC} , in which the TCC is inverted so that a value of θ can be estimated directly from X and (b) the expected *a posteriori*—or Bayesian posterior mean—estimates, θ_{EAP} . Using data from Tenth-Grade English and Math Tests, the conditional, expected values for θ_{TCC} and θ_{EAP} (using both normal $N(0, 1)$ and $N(0, 10)$ priors), along with their conditional standard errors, were computed and plotted against a grid of actual θ s. Under a normal $N(0, 1)$ prior, it was found that the Bayesian θ_{EAP} s showed considerably smaller standard errors of measurement compared with the θ_{TCC} s—especially in the tails of the θ -distribution. However, the bias of the θ_{EAP} s based on the $N(0, 1)$ prior was substantial in the extremes of the distribution of θ . The normal $N(0, 10)$ prior for computing the θ_{EAP} s reduced their bias but increased their standard error—These were not unexpected statistical results, given the nearly universal trade-off between bias and standard error. The choice among the three summed-score θ -estimates examined here depends largely on which of the two major sources of distortion—bias versus standard error—is the more harmful.

Keywords

achievement testing, Bayesian, estimation, item response theory, reliability

Introduction

In certain areas of measurement—especially educational measurement—it is required that identical values of a summed, number-correct score be assigned identical values of an item response theory (IRT) θ -estimate—as opposed to determining θ -estimates based on an individual's entire vector of responses to all test items. In situations where the same number-correct score must be assigned the same θ -estimate, it is not uncommon to find IRT scoring in which the test characteristic curve (TCC) is inverted to produce a table of values of X along with their corresponding θ -estimates. This type of θ -estimation is fairly common in the scoring of tests used in No Child

¹Pacific Metrics Corporation, Monterey, CA, USA

Corresponding Author:

W. Alan Nicewander, Pacific Metrics Corporation, 1 Lower Ragsdale Drive, Building 1, Suite 150, Monterey, CA 93940, USA.

Email: anicewander@pacificmetrics.com

Left Behind programs. The points on the TCC are expected values of X given θ , $E(X|\theta)$, and psychometrically, are the true scores for X given θ . Statistically, the TCC is the regression function for the number-correct score on θ . In IRT score-estimation, this regression function is “inverted” to produce a table of expected X s and their corresponding θ values—let this inverted regression function be symbolized as $E^{-1}(X|\theta)$. The TCC-based estimate is discussed by (Yen, 1984), who refers to the method as a “first-order approximation” to a maximum likelihood solution for θ . From the point of view of psychometrics, these estimates are problematical in that observed scores are treated as though they were true scores in their computation. Statistically, these estimates are problematical in that they are based on the “wrong regression function,” $E(X|\theta)$. If one is seeking to estimate θ from X , then it is arguable that the regression function should be $E(\theta|X)$, the regression of θ on X , or equivalently, the mean of the posterior distribution of θ , given X . Bayesian expected a posteriori (EAP) θ -estimates are found by solving for $E(\theta|X)$ for each value of X using some specified prior distribution of θ (see Thissen & Orlando, 2001). The primary reason for choosing a Bayesian X -to- θ scoring procedure is that the resulting θ -estimates should have lower standard errors of measurement (SEMs) compared with those based on the TCC, and that the statistical and psychometric foundations for these estimates are more sound than for the inverted regression, $E^{-1}(X|\theta)$, that is used in estimates based on the TCC. Because Bayesian EAP estimates are the means of *posterior* distributions of θ , some *prior* distribution of θ must be specified. In this inquiry, normal $N(0, 1)$ and $N(0, 10)$ distributions were used as priors. First, consider the nature of the two IRT scoring methods—one based on $E^{-1}(X|\theta)$, and the other based on the regression function, $E(\theta|X)$. In what follows, n is the maximum value of X , and $f_{icc}(X)$ and $f_{cap}(X)$ represent the two IRT transformations of X to θ -estimates.

θ_{TCC} Estimates

The typical method for computing θ_{TCC} s is to linearly interpolate values of θ from the inverted TCC. Instead of interpolating the values of θ corresponding to each number-correct score, X , it was convenient mathematically to use nonlinear least squares for fitting three-parameter logistic (3PL) curves to the TCCs—to yield a parameterized TCC—and then directly solving algebraically for the values of θ . In this way, the θ_{TCC} s could be expressed as direct functions of X , the number-correct score, without the necessity of interpolating values from a TCC. This solution for the θ_{TCC} s begins with the proportion-correct TCC expressed as a 3PL function,

$$X/n = \gamma + \frac{(1 - \gamma)}{1 + \exp(-\alpha(\theta - \beta))}, \quad (1)$$

where X is the number-correct score, and n is either the number of items on the test or the maximum number of points—depending where a test is composed dichotomous, multiple-choice (MC) items or a mix of dichotomous and polytomous (CR) items. Then, using SAS® PROC NLIN, least squares estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$ were computed, and using these estimates, Equation 1 was solved for θ .

$$\theta = \theta_{TCC} = f_{icc}(X) = \frac{1}{-\hat{\alpha}} \ln \left(\frac{1 - (X/n)}{(\hat{\gamma} - (X/n))} \right) + \hat{\beta}. \quad (2)$$

In applying Equation 2 to the sample data used here, values of X/n equal to 1 were set to .99, and values of $X/n \leq \hat{\gamma}$ were set equal to $\hat{\gamma} + .001$; in this way, Equation 2 always had a solution regardless of the value of X . Estimates were further constrained to be in the interval

(−5, 5). These “fix-ups” at the extremes of the X -distribution are heuristic for the data analyzed here and are discussed in the “Conclusion” section. The PROC NLIN SAS code was applied to a grid of 41 values of θ (−4, 4 by steps of .2). Closeness of fit was summarized by computing (a) nonlinear $R^2 = SS_{model} / SS_{total}$ and (b) plots of the actual (proportion-correct) TCCs and the 3PL-fitted versions of the same. The actual TCCs and estimated TCCs were essentially coincident in the present study. The SAS code for the nonlinear regression estimates of the TCC parameters in Equation 1 is extremely simple and is shown in the appendix along with graphs showing the concordance between the actual TCCs and approximate TCCs based on fitted 3PL curves. Also given on the graphs are the 3PL parameter estimates for the TCCs. The experience with fitting TCCs with 3PL functions has shown similar close degrees of fit between actual and 3PL-fitted TCCs.

For fixed values of actual θ , the conditional means and conditional variances for the θ_{TCC} s were computed as,

$$E(\theta_{TCC}|\theta) = \sum_{X=0}^n f_{tcc}(X)p(X|\theta) \quad (3)$$

and

$$\sigma^2[\theta_{TCC}|\theta] = \sum_{X=0}^n f_{tcc}^2(X)p(X|\theta) - [E(\theta_{TCC}|\theta)]^2. \quad (4)$$

The conditional probabilities, $p(X|\theta)$, follow the compound binomial law (see Lord & Novick, 1968). In this inquiry, the values of $p(X|\theta)$ were computed using Hanson's (1994) extension of the Lord and Wingersky (1984) recursion formula that allows the use of dichotomous and/or polytomous items calibrated using IRT models (see Kolen & Brennan, 2004).

θ_{EAP} Estimates

These estimates were computed using

$$\theta_{EAP} = E(\theta|X) = f_{eap}(X) = \frac{1}{p(X)} \int_{-\infty}^{+\infty} p(X|\theta)\varphi(\theta)\theta d\theta. \quad (5)$$

where $\varphi(\theta)$ was either the normal probability density $N(0, 1)$ or $N(0, 10)$, and $p(X) = \int_{-\infty}^{+\infty} p(X|\theta)\varphi(\theta)d\theta$ is the marginal probability of the score, X . The integration indicated in Equation 5—and for $p(X)$ —was done numerically using 41 quadrature values of θ (−4, 4 in steps of .2). Re-normed, normal probability densities were used as quadrature weights that summed to unity. For these fixed values of θ , the conditional means and conditional variances for the $\theta_{EAP-N(0,1)}$ estimates were computed as

$$E(\theta_{EAP}|\theta) = \sum_{X=0}^n f_{eap}(X)p(X|\theta), \quad (6)$$

and

$$\sigma^2(\theta_{EAP}|\theta) = \sum_{X=0}^n f_{eap}^2(X)p(X|\theta) - [E(\theta_{EAP}|\theta)]^2. \quad (7)$$

Reliability Coefficients for θ_{EAP} and θ_{TCC}

Let θ_{EST} represent θ_{TCC} , $\theta_{EAP-N(0,1)}$, or $\theta_{EAP-N(0,10)}$. Reliability coefficients for Tenth-Grade English (ELA-10) and Tenth-Grade Math Tests (Math-10) were computed as the ratio of true to true-plus-error variance as follows:

1. Let $\tau(\theta_{EST}|\theta) = E(\theta_{EST}|\theta)$ = true score for θ_{EST} at θ .
2. Let $E[\tau(\theta_{EST})] = \int_{-\infty}^{+\infty} \tau(\theta_{EST}|\theta)\varphi(\theta)d\theta$ = marginal mean (or average true score) for θ_{EST} .
3. Let $\sigma^2[\tau(\theta_{EST})] = \int_{-\infty}^{+\infty} [E(\theta_{EST}|\theta)]^2\varphi(\theta)d\theta - \{E[\tau(\theta_{EST})]\}^2$ = marginal, true variance for θ_{EST} .
4. Let $\sigma^2[\varepsilon(\theta_{EST})] = \int_{-\infty}^{+\infty} \sigma^2(\theta_{EST}|\theta)\varphi(\theta)d\theta$ = marginal, variance of measurement errors for θ_{EST} .

Finally, $\rho(\theta_{EST})$, the reliability coefficient for θ_{EST} is given by,

$$\rho(\theta_{EST}) = \frac{\sigma^2[\tau(\theta_{EST})]}{\sigma^2[\tau(\theta_{EST})] + \sigma^2[\varepsilon(\theta_{EST})]}. \quad (9)$$

A Small Study Investigating θ_{TCC} and θ_{EAP} Scoring Applied to Math-10 and ELA-10 Tests

ELA-10 Test

This test was composed of 57 items—33 MC items, 20 two-category CR items, and 4 five-category CR items; total points = 69 (mean score = 64% of maximum score). Figure 1 displays the information function for the X , the number-correct score, and in Figure 2, the information function, $I(X, \theta)$ (see Lord, 1980), has been converted to a graph of conditional reliability coefficients, $\rho_\theta = I(X, \theta)/[1 + I(X, \theta)]$ (see Nicewander, 2005; Nicewander & Thomasson, 1999). It has been found convenient to use conditional reliability coefficients as a way of mapping the (unbounded) information values into the unit interval (0, 1)—thereby, making it somewhat easier to summarize the degree of measurement precision. The ELA-10 Test measures more precisely in the lower regions of the ability distribution, as witnessed by conditional reliabilities of .80 or higher, roughly between -3 and 1.5 on the θ -scale. The test did not measure precisely for abilities 1.5 and higher and could be characterized as *easy* (because the M score = 64% correct). Figure 3 contains the plot of the expected θ_{TCC} s and θ_{EAP} s against true θ for the ELA-10 Test. The inward bias (or shrinkage) of the $\theta_{EAP-N(0,1)}$ s, based on a normal $N(0, 1)$ prior, is evident in contrast to the θ_{TCC} s, which were essentially unbiased, and the $\theta_{EAP-N(0,10)}$ s that were intermediate in their bias.

Most of the bias in the $\theta_{EAP-N(0,1)}$ s was removed by using the $N(0, 10)$ prior. The conditional SEMs for the scoring methods are summarized in Figure 4. The striking features about Figure 4 are how large the θ_{TCC} SEMs are in the extremes of the θ -distribution and how uniformly small the $\theta_{EAP-N(0,1)}$ SEMs are. The $\theta_{EAP-N(0,10)}$ s yielded SEMs nearly identical to those of the θ_{TCC} s

Figure 1. ELA-10 information function (number-correct score).
Note. ELA-10 = 10th-Grade English Test.

Figure 2. ELA-10 conditional reliability coefficients.
Note. ELA-10 = 10th-Grade English Test.

for the major part of the θ -distribution $(-3, 3)$; however, the $\theta_{EAP_{J1}(0, 10)}$ SEMs were much smaller than those for the θ_{TCCS} in the extremes of the ability distribution.

Math-10 Test

This test was composed of 64 items: Sixty MC items and 4 five-category CR items for a total of 76 points. It was more difficult than the ELA-10 Test with a mean score equal to 49% of maximum score. The information function for the Math-10 Test is presented in Figure 5, and the information function transformed to conditional reliabilities is displayed in Figure 6. In contrast to the ELA-10 Test, Math-10 measured with most precision in the upper portion of the ability distribution. The conditional reliabilities for Math-10 were at least .80 for a range of abilities

Figure 3. ELA-10 expected TCC- θ s and EAP- θ s (computed using normal priors, $N(0, 1)$ and $N(0, 10)$).
 Note. ELA-10 = 10th-Grade English Test; TCC = test characteristic curve; EAP = expected a posteriori.

Figure 4. ELA-10 conditional SEMs for TCC- θ s and EAP- θ s (computed using normal priors, $N(0, 1)$ and $N(0, 10)$).
 Note. ELA-10 = 10th-Grade English Test; SEM = standard error of measurement; TCC = test characteristic curve; EAP = expected a posteriori.

that included the interval $-1 \leq \theta \leq 4$. The conditional means for the two estimates for the Math-10 Test are given in Figure 7. As was the case for the ELA-10 Test, the θ_{TCC} s were relatively unbiased compared with the $\theta_{EAP_{N(0,1)}}$ s—which showed considerable shrinkage (even more so than was the case for the ELA-10 Test). The $\theta_{EAP_{N(0,10)}}$ s were much less biased than those computed using the normal $N(0, 1)$ prior—with a degree of bias about the same degree as the θ_{TCC} s.

The conditional SEMs for the three estimates, applied to the Math-10 Test, are shown in Figure 8. The SEMs were exceedingly large for the θ_{TCC} s for low values of θ , but of reasonable size at the center of the θ -distribution ($-1 \leq \theta \leq 1$). Conditional SEMs for $\theta_{EAP_{N(0,1)}}$ were, as in the case of the ELA-10 Test, relatively small compared with those for the θ_{TCC} estimates

Figure 5. Math-10 information function.

Figure 6. Math-10 conditional reliability coefficients.

and relatively homogeneous across the θ -distribution. The reason for the relatively high SEMs for θ_{TCC} at the bottom of the θ -distribution was that this math test was fairly difficult and did not have enough easy items to measure well at the lower end of the proficiency distribution. High SEMs were evident for the $\theta_{EAP-\mathcal{M}(0,10)}$ estimates in the lower portion of the θ -distribution ($-3 \leq \theta \leq -1$). The SEMs for $\theta_{EAP-\mathcal{M}(0,10)}$ were fairly moderate for the top of the latent distribution ($0 \leq \theta \leq 4$).

Reliability Coefficients for the θ -Estimates

Table 1 summarizes the reliability estimates for the EAP- and TCC-ability estimates. For ELA-10, the $\theta_{EAP-\mathcal{M}(0,1)}$ estimate had reliability slightly larger than the θ_{TCCS} (.844 vs. .840). For ELA-10, the $\theta_{EAP-\mathcal{M}(0,10)}$ estimates had reliability equal to .837. For Math-10, the $\theta_{EAP-\mathcal{M}(0,1)}$ s had higher reliability than the θ_{TCCS} (.904 vs. .831), and the $\theta_{EAP-\mathcal{M}(0,10)}$ estimates had reliability equal to .874. These estimates were all consonant with the reliabilities for the number-correct scores that are also presented in Table 1.

Figure 7. Math-10 expected TCC- θ s and EAP- θ s (computed using normal priors, $N(0, 1)$ and $N(0, 10)$). Note. TCC = test characteristic curve; EAP = expected a posteriori.

Figure 8. Math-10 conditional SEMs for TCC- θ s and EAP- θ s (computed using normal priors, $N(0, 1)$ and $N(0, 10)$). Note. SEM = standard error of measurement; TCC = test characteristic curve; EAP = expected a posteriori.

Table 1. Reliabilities for ELA and Math Tests Under Number-Correct and IRT Scoring Methods.

Test	Number correct	θ_{TCC}	$\theta_{EAP_n(0, 1)}$	$\theta_{EAP_n(0, 10)}$
ELA-10	.848	.840	.844	.837
Math-10	.912	.831	.904	.874

Note. ELA-10 = 10th-Grade English Test; IRT = item response theory; TCC = test characteristic curve; EAP = expected a posteriori.

Summary of Bias and SEM for EAP and TCC θ -Estimates

Bias

1. The bias in the θ_{TCC} estimates was considerably smaller than that of the $\theta_{EAP-\mathcal{N}(0,1)}$ estimates for the two tests, Math-10 and ELA-10. For both tests, the θ_{TCC} s were essentially unbiased with both having virtually the same range, ± 4 for Math-10 and ELA-10.
2. However, for ELA-10, the range of the $\theta_{EAP-\mathcal{N}(0,1)}$ estimates was $(-3 \leq \theta \leq 3)$, thus, showing shrinkage. Using a $N(0, 10)$ prior, the range of the $\theta_{EAP-\mathcal{N}(0,10)}$ estimates was $(-3.5 \leq \theta \leq 3.3)$, indicating somewhat less shrinkage.
3. For Math-10, the range of the $\theta_{EAP-\mathcal{N}(0,1)}$ estimates was $(-2 \leq \theta \leq 2.7)$, indicating considerable bias; whereas, the range of the $\theta_{EAP-\mathcal{N}(0,10)}$ estimates was $(-3.3 \leq \theta \leq 4)$, indicating that the $N(0, 10)$ prior yielded much less shrinkage than the $N(0, 1)$.
4. It is almost certainly the case that in most applications, the findings here with regard to bias will hold; namely, the θ_{TCC} estimates will be essentially unbiased, and the $\theta_{EAP-\mathcal{N}(0,1)}$ will often be biased inward to a considerable degree—with estimates based on uninformative priors, such as the $\theta_{EAP-\mathcal{N}(0,10)}$, being intermediate in terms of bias.
5. What is also apt to generalize from this inquiry is that difficult tests (such as Math-10) are likely to result in $\theta_{EAP-\mathcal{N}(0,1)}$ s with large bias at the lower end of the proficiency distribution—produced by the prior “kicking in” to make up for the test items’ lack of information (about θ) for low levels of proficiency. A similar result would be expected to occur for extremely easy tests (considerably easier than the ELA-10 Test studied here)—namely, the lack of informative items in the upper extreme of the proficiency distribution would cause the $\theta_{EAP-\mathcal{N}(0,1)}$ s to be “swamped” by the prior distribution and regress the estimates toward the mean value of θ .

Conditional Standard Errors

1. For both ELA-10 and Math-10, the conditional SEMs for the $\theta_{EAP-\mathcal{N}(0,1)}$ s were relatively small and very homogeneous in magnitude across the θ -distribution.
2. For ELA-10, the SEMs for $\theta_{EAP-\mathcal{N}(0,1)}$ s and $\theta_{EAP-\mathcal{N}(0,10)}$ s were fairly similar to the θ -interval $(-1, 1)$, but SEMs for $\theta_{EAP-\mathcal{N}(0,10)}$ s were much larger than those for $\theta_{EAP-\mathcal{N}(0,1)}$ s outside this interval.
3. For Math-10, the SEMs for $\theta_{EAP-\mathcal{N}(0,1)}$ s and $\theta_{EAP-\mathcal{N}(0,10)}$ s were similar in the θ -interval $(0, 2)$, but the $\theta_{EAP-\mathcal{N}(0,10)}$ s yielded SEMs that were considerably larger than $\theta_{EAP-\mathcal{N}(0,1)}$ s outside this interval.
4. For Math-10, the θ_{TCC} s had SEMs that were virtually identical to those for the $\theta_{EAP-\mathcal{N}(0,10)}$ s for most of the θ -distribution $(-1 \text{ to } 4)$, but considerably larger than those for $\theta_{EAP-\mathcal{N}(0,1)}$ s for values of θ outside the θ -interval $(-1, 2)$. For values of $\theta \leq -1$, the SEMs for θ_{TCC} s were extremely large.
5. For ELA-10, the θ_{TCC} s had SEMs that were virtually identical to those for the $\theta_{EAP-\mathcal{N}(0,10)}$ s for the majority of the θ -distribution $(-3 \text{ to } 3)$ but notably larger than those for $\theta_{EAP-\mathcal{N}(0,1)}$ s for values of θ outside the θ -interval $(-2, 1)$.

Figure 9. Inverted TCC, $E^{-1}(X|\theta)$, for ELA-10.

Note. TCC = test characteristic curve; ELA-10 = 10th-Grade English Test.

Conclusion

Two regression functions— $E^{-1}(X|\theta)$ and $E(\theta|X)$ —were compared with regard to their estimates of θ (and SEMs) computed from the number-correct score, X . The results of this small study indicated that, at the level of the individual examinee, both $\theta_{EAP-N(0,1)}$ and $\theta_{EAP-N(0,10)}$ estimates had the higher measurement precision—as indicated by their smaller conditional SEMs and larger reliabilities relative to the θ_{TCC} estimates. However, the $\theta_{EAP-N(0,1)}$ estimates had considerable regression toward the mean θ , and the $\theta_{EAP-N(0,10)}$ estimates were a compromise estimate having, for the most part, reasonable degrees of bias and SEMs relative to the θ_{TCC} and $\theta_{EAP-N(0,1)}$ estimates. Assuming that the results of our study have a degree of generalizability, it is clear that the choice of a summed-score θ -estimate will depend largely on which of these two sources of distortion—bias versus standard error—is the more harmful. There may well be situations where bias in a θ -estimate is more serious than measurement error—for example, in cases where a cut-score is at the high-or-low ends of the proficiency distribution. In a general sense, if cut-scores are involved, the bias-SEM trade-off depends on where the cut-scores lie in the ability distribution. If cut-scores are not involved, then, it is the opinion of the present authors that, in most educational measurement situations, a θ -estimate having small and homogeneous measurement errors across the range of abilities would trump bias in terms of desirability. If one agrees with this position, then the clear choice of estimator would be the $EAP\theta$ -estimate based on a $N(0, 1)$ prior. If lack of bias is one's principal criterion for an IRT score, then the θ_{TCC} estimate would seem to be a better choice, except for its very large SEMs in the extremes of the proficiency distribution. Based on the present inquiry, the $EAP\theta$ -estimate, based on a $N(0, 10)$ prior (or some other uninformative prior), constitutes a compromise estimate (even though it is a bit “grubby” statistically) when low bias is primary and small SEMs are secondary.

One can get an idea about how these three scoring methods will operate on a particular test by looking at its information function (or the information function rescaled to be conditional reliabilities—as done here). For example, Figure 2 indicates that the conditional reliabilities for the ELA-10 Test varied only between .60 and .85 over the range of θ s. Therefore, this test measured fairly accurately over a wide range of latent proficiency, and one would expect modest shrinkage and small SEMs in the $\theta_{EAP-N(0,1)}$ estimates—which was the case (see Figures 3 and

Figure 10. Inverted TCC, $E^{-1}(X|\theta)$, for Math-10.
 Note. TCC = test characteristic curve.

4). Arguably, $\theta_{EAP_{\mathcal{H}}(0,1)}$ would be the choice of estimates for ELA-10. For Math-10, the information function, rescaled as conditional reliabilities in Figure 6, tells a different story. This test showed strong reliability (.80s and .90s) over the top half of the θ -range, but Figure 6 clearly implies that the test was deficient in items appropriate for measuring the lower proficiency levels; therefore, $\theta_{EAP_{\mathcal{H}}(0,1)}$ estimates may produce too much shrinkage to be useful for measuring low levels of proficiency. It could be conjectured that $\theta_{EAP_{\mathcal{H}}(0,10)}$ or θ_{TCC} estimates may be preferable to $\theta_{EAP_{\mathcal{H}}(0,1)}$, and this is confirmed by Figure 7, where both $\theta_{EAP_{\mathcal{H}}(0,10)}$ and θ_{TCC} show very low levels of shrinkage. However, Figure 8 implies that θ_{TCC} is probably not acceptable because of very large SEMs in the lower extremes of the θ -distribution. By process of elimination, $\theta_{EAP_{\mathcal{H}}(0,10)}$ emerges as the better IRT score for Math-10.

Some Final Notes

The $\theta_{EAP_{\mathcal{H}}(0,1)}$ estimate is sometimes disparaged because of its large bias in extremes of the θ -distribution—as if this were completely the fault of the estimate. However, the truth is that this bias is partially a property of the *test* in that the estimate is swamped by the prior distribution due to the lack of item data for extreme levels of latent proficiency. Therefore, one way to lessen the shrinkage of the $\theta_{EAP_{\mathcal{H}}(0,1)}$ estimates is to add more easy and/or difficult items to a test. If this is not possible, which is often the case, then switching to a uniform, large-variance normal, or other noninformative prior, is often the only practical way to “unshrink” *EAP* θ -estimates and maintain acceptable levels of SEMs. One will notice in Figures 4 and 8 that the SEMs for both of the θ_{EAPS} increase from the central part of the θ -distribution and then begin to decrease at the extremes of the distribution. The SEMs for θ_{EAPS} are smallest where the items most efficiently measure θ ; then increase as item efficiency decreases—until the prior distribution begins to swamp the item information, resulting in shrinkage of both θ -estimates and SEMs. The mechanism for the shrinkage of the $\theta_{EAP_{\mathcal{H}}(0,1)}$ s—and their SEMs—can be explained in this way: If one looks at the integral in Equation 6 for computing $\theta_{EAP_{\mathcal{H}}(0,1)}$, these estimates can be thought of as weighted sums of the actual θ s, where the weights are the joint probabilities, $p(X|\theta)\varphi(\theta)$. If there are few difficult-and-easy items, then in the tails of the θ -distribution, the extreme values of X have

small values of $p(X|\theta)$, which are made vanishingly small by multiplying them by the low values of the normal densities, $\phi(\theta)$, in the tails of the $N(0, 1)$ θ -distribution. Therefore, one is basically left with $\theta_{EAP-N(0,1)}$ s that place negligible weights on the tails of the θ -distribution, and, as a result, the estimates are based on central values of θ . Therefore, estimates are regressed toward the mean of θ and have shrinking SEMs. This shrinkage in the values and variability of the $\theta_{EAP-N(0,1)}$ estimates is lessened by using a prior distribution—such as a $N(0, 10)$, rectangular, or other noninformative prior distribution—where the tail densities are not vanishingly small.

The major problem with θ_{TCC} s—namely, large SEMs in the extremes of the proficiency distribution—cannot be ameliorated either with prior distributions (because no prior is involved in computing θ_{TCC} s), or with addition of additional easy-and-difficult items. The basic problem with θ_{TCC} s can be seen in the striking graphs of the inverted TCCs for ELA-10 and Math-10 shown in Figures 9 and 10. In these graphs, one can see that the upper and lower asymptotes of the TCCs ($\hat{\gamma}$ and 1.00) cause havoc in $E^{-1}(X|\theta)$ by becoming asymptotes of $\pm \infty$; thus, sending θ_{TCC} s “into outer space” at the upper end of the number-correct scale and “off the edge of a cliff” on the lower end. More to the point, at both ends of the number-correct score distribution, small changes in the score, X , can cause huge changes in the values of θ_{TCC} —These large changes, of course, translate into large SEMs for θ_{TCC} s in the extremes of the θ -distribution. Recall that in the present investigation, “fix-ups” were used at the ends of the number-correct distribution: (a) the proportion-correct score, X/n , was set to $\hat{\gamma} + .001$ for values of $X/n \leq \hat{\gamma}$, the lower asymptote of the 3PL fitted to the TCCs; (b) values of X/n equal to unity were set to .99, before inserting them in Equation 3 to estimate θ_{TCC} ; and (c) θ_{TCC} s were further restricted by constraining them to the interval $(-5, 5)$. These fixes are heuristic (i.e., they gave reasonable values for these data), and other data could require other remedies. If these fix-ups had not been used, the SEMs for the θ_{TCC} s would have been *much* larger than they are in Figures 8 and 9. It is important to note that the decreases in the size of the SEMs for the θ_{TCC} s in the extremes of the θ -distribution are due to the fix-ups that were used—without them, the SEMs would continue to increase. The necessity for these score adjustments for θ_{TCC} s points to another advantage of the θ_{EAP} s—They can be computed for all values of X without the necessity of (somewhat arbitrary) fixes for extreme values of X .

Appendix

SAS Code for Fitting a Logistic Curve to a Proportion-Correct TCC

(Input is a TCC—[i.e., X s and θ s], and n [max. score])

DATA;

INPUT

 x theta n;

 pc_tcc=x/n;

PROC NLIN;

PARMS

 alpha=.5 to 2 by .5

 beta= -2 to 2 by .5

 gamma=0 to .5 by .1;

MODEL

 pc_tcc =gamma+(1-gamma)/(1+exp(-alpha*(theta-beta)));

RUN;

TCCs and Estimated TCCs for Math-10 and ELA-10 Tests

Figure A1. TCC and 3PL-estimated-TCC for ELA-10 test.

Note. TCC = test characteristic curve; 3PL = three-parameter logistic; ELA-10 = 10th-Grade English Test.

Figure A2. TCC and 3PL-estimated-TCC for Math-10 test.

Note. TCC = test characteristic curve; 3PL = three-parameter logistic; ELA-10 = 10th-Grade English Test.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Hanson, B. A. (1994). *Extension of Lord-Wingersky algorithm to computing test score distributions for polytomous items*. Unpublished manuscript. Retrieved from <http://www.b-a-h.com/papers/note9401.pdf>
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating methods and practice*. New York, NY: Springer-Verlag.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score equatings. *Applied Psychological Measurement*, 8, 453-461.
- Nicewander, W. A. (2005, October). *Relationships between the indices of measurement precision from classical test theory and IRT*. Paper presented at the annual meeting of the Society for Multivariate Experimental Psychology, Lake Tahoe, NV.
- Nicewander, W. A., & Thomasson, G. D. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement*, 23, 239-247.
- Thissen, D., & Orlando, M. (2001). IRT for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum.
- Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, 21, 93-111.