

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

Alan Nicewander, Ph.D.
Matthew Schulz, Ph.D.
Howard Mitzel, Ph.D.



Pacific Metrics Corporation
585 Cannery Row, Suite 201
Monterey, California 93940

Background

Several testing organizations use IRT scoring methods that invert the test characteristic curve (TCC) in order to produce a table of number-correct (NC) to θ (and hence, to a scale) tables. Call θ s estimated in the manner, $TCC(\theta)$ s. This method treats an observed, number correct (NC) score as though it were a true score. The regression of the true score on θ is given by the test characteristic curve, and this regression is “inverted” to produce a table of the regression of θ on the NC true scores. The $TCC(\theta)$ estimate is discussed by (Yen, 1984), who refers to the method as a “first-order approximation” to a maximum likelihood solution for θ based on the NC score. The $TCC(\theta)$ estimates are relatively unbiased, but in some cases, have large conditional SEMs-- especially at the lower and upper ends of the θ -distribution.

Perhaps a preferred approach to $TCC(\theta)$ s is the Bayesian, expected *a-posteriori* θ -estimate based on the NC score [EAP(θ)] (See Thissen & Orlando, 2001, p.119). The primary reason for choosing a NC-to-EAP(θ) IRT scoring procedure is that the EAP(θ)’s should have lower SEMs compared to the $TCC(\theta)$ ’s, and that the statistical theory for these estimates is more sound than the inverted regression that is employed by $TCC(\theta)$. Since an EAP(θ) is a mean of a *posterior* distribution of θ , some *prior* distribution of θ must be specified. In this inquiry Normal(0,1) prior distribution was used.

A Small Study Investigating $TCC(\theta)$ and the EAP(θ) Scoring Applied to 10th Grade Math and English Tests

First, consider the nature of the two IRT scoring methods (in what follows, X is the NC score, n is the maximum NC score, and $f_{tcc}(X)$ and $f_{eap}(X)$ represent the two, IRT transformations of X to θ -estimates:

TCC(θ) Estimates. Instead of interpolating the values of true score and θ that make up a TCC, we used nonlinear least squares to fit logistic curves to the TCCs, and then solved for the θ s in a manner described by Lord (1980, P. 60),

$$TCC(\theta) = f_{tcc}(X) = \frac{1}{-a^*} \log\left(\frac{X/n - c^*}{1 - X/n}\right) + b^*, \quad (1)$$

where a^* , b^* and c^* are least squares estimates of the parameters for fitting a three-parameter logistic (3-PL) function to the TCC, and X/n is the proportion-correct score. It should be noted that in our experience, a 3-PL curve can be fitted to a TCC with extraordinary accuracy—e.g., non-linear R^2 s are invariably over .99 and frequently .9999 or above. For fixed values of true θ , the conditional means and conditional variances for the $TCC(\theta)$ s were computed as,

$$E[TCC(\theta) | \theta] = \sum_{X=0}^n f_{tcc}(X) p(X | \theta), \quad (2)$$

and,

$$\sigma^2[TCC(\theta) | \theta] = \sum_{X=0}^n f_{tcc}^2(X) p(X | \theta) - E[TCC(\theta) | \theta]^2, \quad (3)$$

where, $p(X|\theta)$ is the conditional probability of the NC score, X , given θ . These probabilities follow the compound binomial law and are computed using an extension of the Lord-Wingersky (1984) recursion formula which allows the use of constructed response items calibrated with a partial credit IRT model (See Kolen & Brennan, 2004).

EAP(θ) Estimates

$$EAP(\theta) = f_{eap}(X) = \frac{1}{p(X)} \int_{-\infty}^{+\infty} p(X | \theta) \phi(\theta) \theta d\theta, \quad (4)$$

where, $p(X) = \int_{-\infty}^{+\infty} p(X | \theta) \phi(\theta) d\theta$ is the marginal probability of the NC score, X . The integration indicated in (4)—and for $p(X)$ —is done numerically using quasi-Gaussian quadrature (in which re-normed, normal probability densities are used as the quadrature weights). For fixed values of θ , the conditional means and conditional variances for the EAP(θ) estimates are computed as.

$$E[EAP(\theta) | \theta] = \sum_{X=0}^n f_{eap}(X) p(X | \theta), \quad (5)$$

and,

$$\sigma^2[EAP(\theta) | \theta] = \sum_{X=0}^n f_{eap}^2(X) p(X | \theta) - E[EAP(\theta) | \theta]^2. \quad (6)$$

Results

10th Grade English

This test was composed of 47 items—22 MC items, 20 3-category CR items and five 5-category CR items. Fig. 1 presents the mean EAP(θ)s and TCC(θ)s for this test. In the calculation of the EAP(θ)s and TCC(θ)s, the prior distribution for θ was set to be Normal (0,1). Fig. 1 contains the plot of the expected TCC(θ)s and EAP(θ)s against true θ . The inward bias of the EAP(θ)s, relative to the TCC(θ) estimates is evident in contrast to the TCC(θ)s which are essentially unbiased. The conditional SEMs for the two scoring methods are summarized in Fig. 2. The striking features about Fig. 2 are how large the TCC(θ) SEMs are in the extremes of the θ -distribution, and how uniformly small the EAP(θ) SEMs are.

10th Grade Math Test

This test was composed of 64 items: 60 MC items and four 5-category CR items. The conditional means for the two estimates for math test are given in Fig. 3. As was the case for English, the TCC(θ)s are relatively unbiased compared to the EAP(θ)s—which show considerable inward bias (even more so than was the case for English).

The conditional SEMs for the two methods applied to the 10th grade math test are shown in Fig. 4. The SEMs are exceedingly large for the TCC(θ)s for low values of θ , but of reasonable size at the center and top of the θ -distribution. The reason for the relatively high SEMs for TCC(θ) at the bottom of the θ -distribution is that this math test fairly difficult and does not measure well at the lower end of the proficiency distribution—and these estimates are not constrained by a prior distribution of θ .

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

Bias--The bias in the TCC(θ) estimates was considerably smaller than that of the EAP(θ) estimates for the two tests, Math and English, Grade 10. For both tests, the true-and-TCC(θ)s had virtually the same range, ± 4 . However, for English, the range of the EAP(θ)s, using a Normal prior, was -3 to +2.5. For Math, the range of the EAP(θ)s was -2 to +3.7

Conditional Standard Errors—Under the assumption of a Normal(0,1) prior, the SEMs for the EAP(θ) estimates were considerably smaller than that of the TCC(θ)s at the extremes of the θ -distribution. In the central part of the θ -distribution, the SEMs for the two methods were quite similar. For Math, the TCC(θ)s had SEMs that exceeded unity at the lower end of the θ -distribution, and for English, the TCC(θ)s had SEMs that were greater than one at the upper end of the θ -distribution

Reliability Coefficients for EAP(θ) and TCC(θ)

For both EAP(θ) and TCC(θ), reliability coefficients for English and Math were computed by integrating the (squared)conditional SEMs across a N(0,1) θ in order to obtain a marginal error variance. The marginal true variances were computed by integrating the squares of the expected θ -estimates across θ , and subtracting the square of the marginal expected value. The reliability coefficients—the ratios of the marginal true-variances to the sum of the marginal true-and-error variances—for the θ -estimates are given in the Table 1 below:

For English, the EAP (Normal prior) has reliability slightly larger than TCC(θ) scoring. For Math10, the EAP(θ)s have considerably higher reliability than the TCC(θ)s.

Conclusions

The results of this small study indicate that, at the level of the individual examinee, the EAP(θ) estimates will have the higher measurement precision—as indicated by their small, conditional SEMs and larger reliabilities relative to the other θ -estimates. However, these estimates have considerable regression inwards toward the mean—and more students will be placed in the middle achievement levels and fewer in the lower and upper levels [compared to TCC(θ) estimates].

References

- Kolen, M. J., Brennan, R. L., (2004). *Test equating methods and practice* (2004). New York: Springer-Verlag
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Lord, F.M. & Wingersky, M. S. (1984) Comparison of IRT true-score and equipercentile observed-score equatings. *Applied Psychological Measurement*, 8, 453-461.
- Thiessen, D. & Orlando, M. (2001). IRT for items scored in two categories. In D. Thiessen & H. Wainer (Eds.), *Test scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates.
- 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.

Table 1. Reliabilities for English and Math tests under two IRT scoring methods

	TCC(θ)	EAP(θ)
English	.837	.844
Math	.826	.913

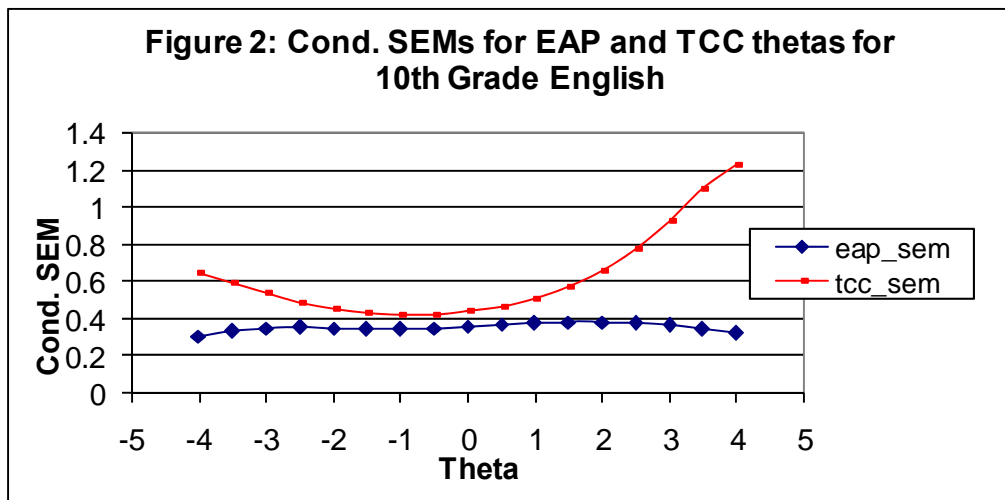
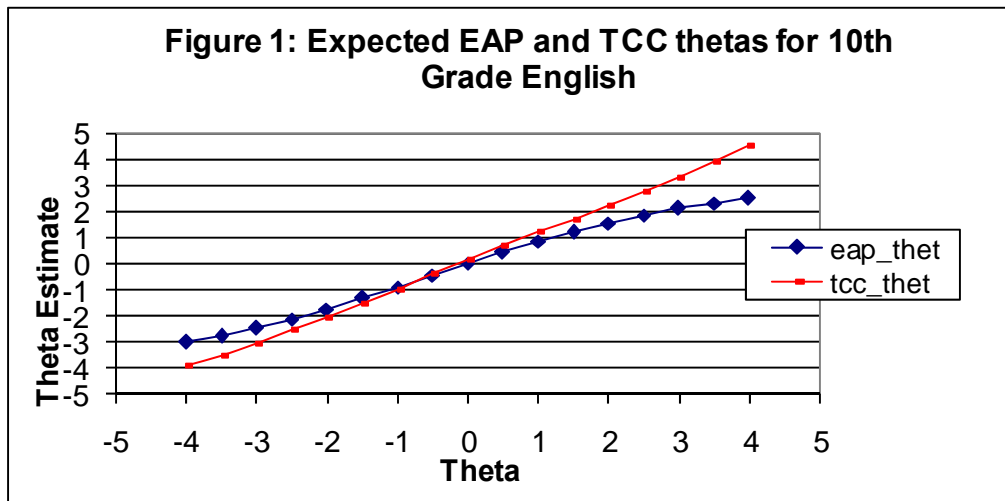


Figure 3: Expected EAP and TCC thetas for 10th Grade Math

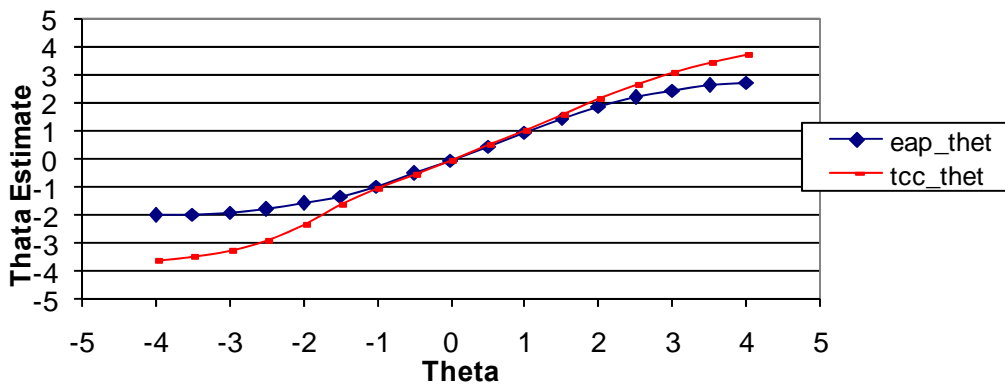


Figure 4: Cond. SEMs for EAP and TCC thetas for 10th Grade Math

