Research Report

Reliability of Scaled Scores

Shelby J. Haberman

December 2008

ETS RR-08-70

Reliability of Scaled Scores

Shelby J. Haberman
ETS, Princeton, NJ

December 2008

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

The reliability of a scaled score can be computed by use of item response theory. Estimated reliability can be obtained even if the item response model selected is not valid.

Key words: Item response theory, variance of measurement, Cronbach α, item sets, generalized partial credit model, two-parameter logistic model

Acknowledgments

The staff in Statistical Analysis assisted with access to data. Neil Dorans provided background on the current computational methods used at ETS and reviewed the manuscript. Helpful comments were also provided by Yi-Hsuan Lee and Dan Eignor.

Reliability of scaled scores can be determined by use of item response theory (Kolen, Zeng, & Hanson, 1996). This approach involves some risk, for the item response model employed may well not hold. An alternative approach based on item response theory does not assume that the model is true (Haberman, 2007). If the model is true, then the two approaches should give very similar results in large samples, so that an indication of the impact of model error is provided. In Section 1, conventional computation of reliability of a scaled score is reviewed. In Section 2, the proposed method of reliability computation is considered. In Section 3, some examples of computations are provided for operational tests. Implications are explored in Section 4.

Throughout the paper, it is assumed that examinee responses X_{ij}, $1 \le j \le q$, $1 \le i \le n$, are available, where examinee i, $1 \le i \le n$, has response X_{ij} to item j, $1 \le j \le q$. It is assumed that q is at least 2. The response vectors $\mathbf{X}_i$ with coordinates X_{ij}, $1 \le j \le q$, are assumed to be independent and identically distributed. Each response X_{ij} is in a finite set A_j of real values, where A_j has at least two elements. The possible values of $\mathbf{X}_i$ may be denoted by A. In many applications, $A_j = \{0, 1\}$, $X_{ij} = 1$ if the response is correct, and $X_{ij} = 0$, otherwise. In formula scoring with a_j multiple choices, one might have $A_j = \{-1/((a_j - 1), 0, 1)\}$ with $X_{ij} = 1$ for a correct response, $X_{ij} = 0$ for an omitted response, and $X_{ij} = -1/(a_j - 1)$ for an incorrect response. For a q-dimensional vector $\mathbf{y}$, let $\Sigma(\mathbf{y})$ be the sum of the coordinates of $\mathbf{y}$. Then the sum $S_i = \Sigma(\mathbf{X}_i)$ of the X_{ij}, $1 \le j \le q$, is the raw score for examinee i. The finite set of possible raw scores is denoted by $\Sigma(A)$. To any possible raw score s in $\Sigma(A)$ corresponds a real scaled score $U(s)$. A one-dimensional item-response model is considered for the responses $\mathbf{X}_i$, $1 \le i \le n$. Under the model, it is assumed that for some K-dimensional vector $\boldsymbol{\beta}$ in a set B with a nonempty interior and for some family of probability distributions $P(\boldsymbol{\gamma})$, $\boldsymbol{\gamma}$ in B, any q-dimensional vector $\mathbf{x}$ in A with coordinate x_j, $1 \le j \le q$, the probability $p(\mathbf{x})$ that $\mathbf{X}_i = \mathbf{x}$ is the expected value $p_*(\mathbf{x}; \boldsymbol{\beta})$ of

$$p_*(\mathbf{x}|\theta; \boldsymbol{\beta}) = \prod_{j=1}^{q} p_j(x_j|\theta; \boldsymbol{\beta}),$$

where $p_j(x_j|\theta; \boldsymbol{\beta}) > 0$, the sum of the $p_j(x|\theta; \boldsymbol{\beta})$, x in A_j, is equal to 1, and θ has a probability distribution $P(\boldsymbol{\beta})$. Thus for a random variable θ_i, the X_{ij}, $1 \le j \le q$, are conditionally independent given θ_i, θ_i has distribution $P(\boldsymbol{\beta})$, and the conditional probability that $X_{ij} = x_j$ given $\theta_i = \theta$ is $p_j(x_j|\theta; \boldsymbol{\beta})$. If defined, the expectation of a real function $g(\theta_i)$ of θ_i is denoted by $E(g(\theta); \boldsymbol{\beta})$. It is assumed that $p_j(x_j|\theta; \boldsymbol{\beta})$ is continuous in both θ and $\boldsymbol{\beta}$. The probability $p_S(s)$, s

in $\Sigma(A)$, that the raw score $S_i = s$ is the sum of the probabilities $p(\mathbf{x})$ for response combinations $\mathbf{x}$ in A such that the sum $\Sigma(\mathbf{x}) = s$. Thus the model assumes that the probability $p_S(s) = p_{S*}(s;\boldsymbol{\beta})$, the corresponding sum of the probabilities $p_*(\mathbf{x};\boldsymbol{\beta})$ for response combinations $\mathbf{x}$ in A such that the sum $\Sigma(\mathbf{x}) = s$. Computation of $p_{S*}(s;\boldsymbol{\beta})$ is not generally difficult if appropriate recursive algorithms are used (Kolen et al., 1996).

Even if the model does not hold, it is assumed that a unique $\boldsymbol{\beta}_*$ exists that minimizes the information measure $E(-\log p_*(\mathbf{X}_1, \boldsymbol{\beta}_*))$ over B (Haberman, 2007). If the model does hold, then $\boldsymbol{\beta}_* = \boldsymbol{\beta}$. The parameter $\boldsymbol{\beta}_*$ can be regarded as the value in B that leads to $p_*(\mathbf{x};\boldsymbol{\beta}_*)$ that best approximate $p(\mathbf{x})$ for $\mathbf{x}$ in A.

In practice, $\boldsymbol{\beta}_*$ is estimated from the $\mathbf{X}_i$ by maximum likelihood. The maximum-likelihood estimate of $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}$. It is assumed that the customary results hold that $\hat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}_*$ with probability 1 and $P(\hat{\boldsymbol{\beta}}_*)$ converges weakly to $P(\boldsymbol{\beta}_*)$ with probability 1. It then follows that $E(g(\theta);\hat{\boldsymbol{\beta}})$ converges with probability 1 to $E(g(\theta);\boldsymbol{\beta})$ if g is bounded and continuous.

1 Computation of Reliability of the Scaled Score

If the model holds, then the scaled score $U_i = U(S_i)$ has conditional variance given $\theta_i = \theta$ (conditional variance of measurement at θ) of

$$\sigma^2(U|\theta;\boldsymbol{\beta}) = \sum_{s\in\Sigma(A)} [U(s) - \mu(U|\theta;\boldsymbol{\beta})]^2 p_{S*}(s|\theta;\boldsymbol{\beta}),$$

where the conditional scale score mean given $\theta_i = \theta$ is

$$\mu(U|\theta;\boldsymbol{\beta}) = \sum_{s\in\Sigma(A)} U(s) p_{S*}(s|\theta;\boldsymbol{\beta}).$$

The expected conditional variance given $\theta_i = \theta$ (variance of measurement) is then $E(\sigma^2(U|\theta;\boldsymbol{\beta});\boldsymbol{\beta})$. The conditional standard error of measurement is the square root of the conditional variance of measurement. If the expected value of $\mu(U|\theta_i;\boldsymbol{\beta})$ is denoted by

$$\mu(U;\boldsymbol{\beta}) = E(\mu(U|\theta;\boldsymbol{\beta});\boldsymbol{\beta}),$$

then the variance of $\mu(U|\theta_i;\boldsymbol{\beta})$ is

$$\sigma^2(\mu(U|\theta;\boldsymbol{\beta});\boldsymbol{\beta}) = E([\mu(U|\theta;\boldsymbol{\beta}) - \mu(U;\boldsymbol{\beta})]^2;\boldsymbol{\beta}).$$

Thus the reliability of the scaled score is

$$\rho^2(U;\boldsymbol{\beta}) = \frac{\sigma^2(\mu(U|\theta;\boldsymbol{\beta});\boldsymbol{\beta})}{\sigma^2(U|\theta;\boldsymbol{\beta}) + \sigma^2(\mu(U|\theta;\boldsymbol{\beta});\boldsymbol{\beta})}$$

(Kolen et al., 1996). Because the model is assumed to hold, $\mu(U;\boldsymbol{\beta})$ is the expected scaled score $E(U_i)$, and the variance $\sigma^2(U_i)$ of the scaled score is $\sigma^2(U|\theta;\boldsymbol{\beta}) + \sigma^2(\mu(U|\theta;\boldsymbol{\beta});\boldsymbol{\beta})$. The maximum-likelihood estimate $E(\sigma^2(U|\theta;\hat{\boldsymbol{\beta}});\hat{\boldsymbol{\beta}})$ of the variance of measurement converges with probability 1 to the variance of measurement $E(\sigma^2(U|\theta;\boldsymbol{\beta});\boldsymbol{\beta})$. The maximum-likelihood estimate $\rho^2(U;\hat{\boldsymbol{\beta}})$ of the reliability converges to the actual reliability $\rho^2(U;\boldsymbol{\beta})$ with probability 1. If $P(\boldsymbol{\beta})$ is the distribution of a polytomous random variable, then the required expectations are readily computed. If $P(\boldsymbol{\beta})$ is the standard normal distribution, as is quite common in item response theory, then Gauss-Hermite quadrature may normally be employed to find expectations.

At ETS, computation of reliability of a scaled score is accomplished by a much older and much cruder approach (Dorans, 1984) based on local linear approximations of scaled scores by raw scores and based on approximation of the distribution of true scores by the empirical distribution of raw scores.

2 Alternative Computation of Reliability of the Scaled Score

An alternative approach to computation of the scaled score does not assume that the model is correct. In this approach (Haberman, 2007), a random variable θ_{i*} has distribution $P(\boldsymbol{\beta})$, and a random vector $\mathbf{X}_{i*}$ with the same possible values as $\mathbf{X}_i$ has conditional probability $p(\mathbf{x}|\theta;\boldsymbol{\beta})$ that $\mathbf{X}_{i*} = \mathbf{x}$ given $\theta_{i*} = \theta$. A random variable θ_i then exists such that the conditional distribution of θ_i given $\mathbf{X}_i$ is the same as the conditional distribution of θ_{i*} given $\mathbf{X}_{i*}$. No assumptions are made concerning the distribution of $\mathbf{X}_i$ other than that each probability $p(\mathbf{X}_i)$ is positive. Let $\mathbf{p}$ denote the function on A with value $p(\mathbf{x})$ at $\mathbf{x}$ in A. If g is a real function on the real line and if the expected value $E(g(\theta);\boldsymbol{\beta}_*)$ of $g(\theta_{i*})$ is defined, then the expected value of $g(\theta_i)$ is the expected value

$$E_*(g(\theta);\mathbf{p},\boldsymbol{\beta}_*) = E(c(\theta;\mathbf{p},\boldsymbol{\beta}_*)g(\theta);\boldsymbol{\beta}_*)$$

of the product $c(\theta_{i*};\mathbf{p},\boldsymbol{\beta}_*)g(\theta_{i*})$. Here the multiplier

$$c(\theta;\mathbf{p},\boldsymbol{\beta}_*) = \sum_{\mathbf{x}\in A} d(\theta;\mathbf{x},\mathbf{p},\boldsymbol{\beta}_*),$$

and the summand

$$d(\theta; \mathbf{x}, \mathbf{p}, \boldsymbol{\beta}_*) = \frac{p(\mathbf{x}) p_*(\mathbf{x}|\theta; \boldsymbol{\beta}_*)}{p_*(\mathbf{x}; \boldsymbol{\beta}_*)}.$$

The conditional probability that $\mathbf{X}_i = \mathbf{x}$ given $\theta_i = \theta$ is then $d(\theta; \mathbf{x}, \mathbf{p}, \boldsymbol{\beta}_*)/c(\theta; \mathbf{p}, \boldsymbol{\beta}_*)$. If the model is valid, then the summand $d(\theta; \mathbf{p}, \mathbf{x}; \boldsymbol{\beta}_*)$ is the conditional probability $p_*(\mathbf{x}|\theta; \boldsymbol{\beta})$ that $\mathbf{X}_i = \mathbf{x}$ given $\theta_i = \theta$, and the multiplier $c(\theta; \mathbf{p}, \boldsymbol{\beta}_*)$ is 1. Thus the expected value $E_*(g(\theta); \mathbf{p}, \boldsymbol{\beta}_*)$ of $g(\theta_i)$ is the same as the expected value of $g(\theta_{i*})$.

The scaled score U_i has conditional variance given $\theta_i = \theta$ (conditional variance of measurement at θ) of

$$\sigma_*^2(U|\theta; \mathbf{p}, \boldsymbol{\beta}_*) = [c(\theta; \mathbf{p}, \boldsymbol{\beta}_*)]^{-1} \sum_{\mathbf{x} \in A} [U(\Sigma(\mathbf{x})) - \mu_*(U|\theta; \mathbf{p}, \boldsymbol{\beta}_*)]^2 d(\theta; \mathbf{x}, \mathbf{p}, \boldsymbol{\beta}),$$

where the conditional scale score mean given $\theta_i = \theta$ is

$$\mu_*(U|\theta; \boldsymbol{\beta}_*) = \sum_{\mathbf{x} \in A} U(\Sigma(\mathbf{x})) d(\theta|\mathbf{x}, \mathbf{p}, \boldsymbol{\beta}).$$

The expected conditional variance given $\theta_i = \theta$ (variance of measurement) is then $E_*(\sigma^2(U|\theta; \mathbf{p}, \boldsymbol{\beta}_*); \mathbf{p}, \boldsymbol{\beta}_*)$. The square root of the conditional variance of measurement is the conditional standard error of measurement. The expected value of $\mu_*(U|\theta_i; \boldsymbol{\beta}_*)$ is the expected value

$$E(U; \mathbf{p}) = \sum_{\mathbf{x} \in A} U(\Sigma(\mathbf{x})) p(\mathbf{x})$$

of U_i, so that the variance of $\mu_*(U|\theta; \mathbf{p}, \boldsymbol{\beta})$ is

$$\sigma_*^2(\mu_*(U|\theta; \mathbf{p}, \boldsymbol{\beta}_*); \mathbf{p}, \boldsymbol{\beta}_*) = E_*([\mu_*(U|\theta; \mathbf{p}, \boldsymbol{\beta}_*) - E(U; \mathbf{p})]^2; \mathbf{p}, \boldsymbol{\beta}_*).$$

Let

$$\sigma^2(U; \mathbf{p}) = \sum_{\mathbf{x} \in A} [U(\Sigma(\mathbf{x})) - E(U; \mathbf{p})]^2 p(\mathbf{x})$$

denote the variance of U_i. Based on the new variable θ_i, the reliability of the scaled score is

$$\rho_*^2(U; \mathbf{p}, \boldsymbol{\beta}_*) = \frac{\sigma_*^2(\mu_*(U|\theta; \mathbf{p}, \boldsymbol{\beta}_*); \mathbf{p}, \boldsymbol{\beta}_*)}{\sigma^2(U; \mathbf{p})}.$$

If $\bar{\mathbf{p}}$ is the fraction of examinees i, $1 \leq i \leq n$, with $\mathbf{X}_i = \mathbf{x}$ and if $\bar{\mathbf{p}}$ is the function on A with value $\bar{p}(\mathbf{x})$ for $\mathbf{x}$ in A, then the estimated variance $\sigma^2(U; \bar{\mathbf{p}})$ converges with probability 1 to the variance $\sigma^2(U; \mathbf{p})$ of U_i. The estimated variance of measurement $E_*(\sigma^2(U|\theta; \bar{\mathbf{p}}, \hat{\boldsymbol{\beta}}; \bar{\mathbf{p}}, \hat{\boldsymbol{\beta}}))$ converges with probability 1 to the variance of measurement $E_*(\sigma^2(U|\theta; \mathbf{p}, \boldsymbol{\beta}_*); \mathbf{p}, \boldsymbol{\beta}_*)$. It follows

that the estimated reliability $\rho_*^2(U;\bar{\mathbf{p}},\hat{\boldsymbol{\beta}})$ of the scaled score converges to the the actual reliability $\rho_*^2(U,\mathbf{p},\boldsymbol{\beta}_*)$ with probability 1. If the model holds, then $\rho_*^2(U,\bar{p},\hat{\boldsymbol{\beta}})$ and $\rho^2(U,\hat{\boldsymbol{\beta}})$ both converge to the same value.

In almost any real test, the vast preponderance of the $\bar{p}(\mathbf{x})$ are 0, for the number of possible combinations of responses is far larger than the sample size. For any real function h on A, the average $\sum_{\mathbf{x}\in A} h(\mathbf{x})$ is equal to $n^{-1}\sum_{i=1}^n h(\mathbf{X}_i)$. In evaluation of variances and reliability coefficients, it is helpful to observe that

$$c(\theta;\bar{\mathbf{p}},\hat{\boldsymbol{\beta}}) = n^{-1}\sum_{i=1}^n \frac{p_*(\mathbf{X}_i|\theta;\hat{\boldsymbol{\beta}})}{p_*(\mathbf{X}_i;\hat{\boldsymbol{\beta}})},$$

$$\sigma_*^2(U|\theta;\bar{\mathbf{p}};\bar{\mathbf{p}}\hat{\boldsymbol{\beta}}) = [c(\theta;\bar{\mathbf{p}},\hat{\boldsymbol{\beta}})]^{-1}n^{-1}\sum_{i=1}^n [U_i - \mu_*(U|\theta;\bar{\mathbf{p}},\hat{\boldsymbol{\beta}})]^2\frac{p_*(\mathbf{X}_i|\theta;\hat{\boldsymbol{\beta}})}{p_*(\mathbf{X}_i;\hat{\boldsymbol{\beta}})},$$

$$\mu_*(U|\theta;\boldsymbol{\beta}_*) = n^{-1}\sum_{i=1}^n U_i\frac{p_*(\mathbf{X}_i|\theta;\hat{\boldsymbol{\beta}})}{p_*(\mathbf{X}_i;\hat{\boldsymbol{\beta}})},$$

$$E(U;\bar{\mathbf{p}}) = n^{-1}\sum_{i=1}^n U_i,$$

and

$$\sigma^2(U;\bar{\mathbf{p}}) = n^{-1}\sum_{i=1}^n [U_i - E(U;\bar{\mathbf{p}})]^2.$$

3 Examples

To illustrate results, reliability computations were made for the scaled scores reported for two forms associated with an ETS assessment. The first form, Form 1, involved about 2,700 examinees and the second form, Form 2, involved about 3,000 examinees. For each form, two sections, Section A and Section B, are considered. For each section, the responses of the examinee for that section are used to construct a raw score total for the section, and the raw score is then converted to a reported scale score for the section. Two approaches were considered. In the first approach, a two-parameter logistic (2PL) model was employed for dichotomous responses with possible values 0 or 1, and a generalized partial credit model was used for polytomous responses. Because both Section A and Section B involved item sets, a second approach was considered in which a generalized partial credit model was applied to the raw score subtotals for each item set. In addition to the IRT analysis, Cronbach α statistics were computed for each total raw score for each section. Two methods are were to compute the Cronbach α. The first method

based computations on individual item responses. The second approach based computations on raw subtotals for each set of items. Results are provided in Tables 1, 2, and 3. For comparison, note that output from statistical analysis performed during equating of Form 2 yielded reliability estimates of 0.871 for the scaled score for Section A on Form 2 and a reliability estimate of 0.888 for the raw score for Section A. For Form 2, the estimated reliability for Section B was 0.870 for the scaled score and 0.872 for the raw score. As previously indicated, these computations are based in Dorans (1984).

Table 1
Reliability Estimates

Estimate	Score	Basis	Form 1		Form 2	
			Section A	Section B	Section A	Section B
Model assumed	Scale	Items	0.907	0.896	0.892	0.876
Model not assumed	Scale	Items	0.904	0.896	0.874	0.862
Model assumed	Raw	Items	0.909	0.898	0.892	0.882
Model not assumed	Raw	Items	0.905	0.898	0.887	0.871
Cronbach α	Raw	Items	0.893	0.892	0.873	0.863
Model assumed	Scale	Sets	0.880	0.877	0.863	0.858
Model not assumed	Scale	Sets	0.880	0.879	0.859	0.849
Model assumed	Raw	Sets	0.882	0.879	0.869	0.865
Model not assumed	Raw	Sets	0.881	0.881	0.869	0.857
Cronbach α	Raw	Sets	0.878	0.870	0.865	0.839

Table 2
Standard Errors of Measurement

Estimate	Score	Basis	Form 1		Form 2	
			Section A	Section B	Section A	Section B
Model assumed	Scale	Items	2.156	2.239	2.191	2.179
Model not assumed	Scale	Items	2.165	2.241	2.214	2.207
Model assumed	Raw	Items	2.763	2.349	2.874	2.041
Model not assumed	Raw	Items	2.771	2.344	2.881	2.040
Cronbach α	Raw	Items	2.944	2.413	3.049	2.108
Model assumed	Scale	Sets	2.423	2.416	2.338	2.287
Model not assumed	Scale	Sets	2.420	2.415	2.336	2.310
Model assumed	Raw	Sets	3.105	2.539	3.097	2.144
Model not assumed	Raw	Sets	3.100	2.531	3.092	2.147
Cronbach α	Raw	Sets	3.142	2.653	3.142	2.285

Table 3
Standard Deviations

Estimate	Score	Basis	Form 1		Form 2	
			Section A	Section B	Section A	Section B
Model assumed	Scale	Items	7.070	6.929	6.659	6.179
model not assumed	Scale	Items	6.985	6.950	6.229	5.937
Model assumed	Raw	Items	9.145	7.352	8.752	5.951
Model not assumed	Raw	Items	9.003	7.349	8.554	5.686
Cronbach α	Raw	Items	9.003	7.349	8.554	5.686
Model assumed	Scale	Sets	7.007	6.894	6.613	6.072
Model not assumed	Scale	Sets	6.985	6.950	6.229	5.937
Model assumed	Raw	Sets	9.037	7.302	8.546	5.829
Model not assumed	Raw	Sets	9.003	7.349	8.554	5.686
Cronbach α	Raw	Sets	9.003	7.349	8.554	5.686

The various estimates are not dramatically different, but differences are still notable. The key issue appears to be the treatment of item sets. Given the same treatment of sets, estimated standard errors of measurement are quite similar for both IRT approaches. As should be expected, the Cronbach α statistics give larger estimated standard errors of measurement than do the corresponding IRT procedures. The IRT estimates are from 1 to 6% smaller. The issue of item sets is somewhat more notable. Set-based estimates are from 3 to 12% larger than are the corresponding item-based estimates. The estimated standard deviations of scores for the method of section 1 in which the model is assumed to hold are often quite close to those for the method of section 2 in which the model is not assumed to hold, but differences can exist. For example, consider the scaled score for Section B of Form 2 for the item-based estimate. The estimate that assumes model validity is about 7% larger than is the estimate that does not assume model validity. The reliability estimates are rather similar for different methods that provide the same treatment or lack of treatment of item sets. Conditional on method of estimate, including treatment of item sets, the reliability results for scale scores and raw scores are rather similar despite some nonlinearity of the raw-to-scale conversion in Form 2 and despite use of rounded scale scores. Effects of item sets are of some concern, especially if one considers percentage changes in terms of differences from 1. For example, in the case of the Cronbach α, for Section B of Form 2, the item-based result is about 15% closer to 1 than is the set-based result.

4 Conclusions

It is quite feasible to estimate reliability of scaled scores with far fewer approximations than are currently used in ETS estimation procedures. At least for the cases examined, the effect of more accurate estimation is not dramatic but not negligible.

The approach to item sets in the analysis is not necessarily the best one. Especially in Section A, in which item sets are quite large, it may be more appropriate to apply restrictive models on the parameters in the generalized partial credit model to overcome concerns about the small numbers of examinees with certain extreme raw subtotals. In addition, it is possible to consider testlet models for the treatment of item sets. The latter choice was avoided in this investigation due to the somewhat higher computational labor involved. Nonetheless, the role of testlet models does warrant study.

The approach that does not assume model validity is not a perfect solution to invalid models, although comparison with the results that assume and do not assume validity can indicate model deficiencies. Nonetheless, analysis of item-based models was still not entirely successful at revealing set effects. Thus it is not realistic to expect that analysis that does not assume a valid model will inevitably lead to a satisfactory treatment of reliability.

References

Dorans, N. J. (1984). *Approximate IRT formula score and scaled score standard errors of measurement at different ability levels* (Tech. Rep. No. SR-84-118). Princeton, NJ: ETS.

Haberman, S. J. (2007). *The information a test provides on an ability parameter* (Research Rep. No. RR-07-18). Princeton, NJ: ETS.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*, 129–140.