Standardized Conditional SEM: A Case for Conditional Reliability

Nambury S. Raju, Larry R. Price, T.C. Oshima and Michael L. Nering

Applied Psychological Measurement 2007; 31; 169
DOI: 10.1177/0146621606291569

Standardized Conditional *SEM*: A Case for Conditional Reliability

Nambury S. Raju, Illinois Institute of Technology
Larry R. Price, Texas State University–San Marcos
T. C. Oshima, Georgia State University
Michael L. Nering, Measured Progress

An examinee-level (or conditional) reliability is proposed for use in both classical test theory (CTT) and item response theory (IRT). The well-known group-level reliability is shown to be the average of conditional reliabilities of examinees in a group or a population. This relationship is similar to the known relationship between the square of the conditional standard error of measurement (*SEM*) and the square of the group-level *SEM*. The proposed conditional reliability is illustrated with an empirical data set in the CTT and IRT frameworks. *Index terms: conditional standard error of measurement; conditional reliability; classical test theory; item response theory*

In the classical test theory framework, it is assumed that the observed score for an examinee $s(x_s)$ is the sum of the true (t_s) score and error (e_s) score. That is,

$$x_s = t_s + e_s. \tag{1}$$

The expectation (E) and the variance (σ^2) of x_s over replications for examinee s may be expressed, respectively, as

$$E(x_s) = E(t_s) + E(e_s) = t_s, \tag{2}$$

$$\sigma^2_{x_s} = \sigma^2_{e_s}. \tag{3}$$

Equations (2) and (3) assume that the error scores have a zero expectation and that the true score for examinee s remains unchanged over replications (Lord & Novick, 1968). The standard deviation of error scores is commonly referred to as the standard error of measurement (SEM) or

$$\sigma_{e_s} = SEM_s. \tag{4}$$

The SEM_s is also referred to as the conditional (or examinee-level) SEM_s because it could vary from one examinee to the next. These conditional SEM_ss are very helpful in practice in interpreting the closeness between an observed score (x_s) and the underlying, unobserved true score (t_s) or in establishing confidence intervals. There are well-established procedures for estimating conditional SEM_ss in classical test theory (Feldt & Brennan, 1989; Feldt, Steffen, & Gupta, 1985; Qualls-Payne, 1992), as well as in item response theory (Hambleton & Swaminathan, 1985; Lord, 1980; Price, Raju, Lurie, Wilkins, & Zhu, 2004; Price, Raju, Lurie, Wilkins, & Zhu, 2006; Wright

Applied Psychological Measurement, Vol. 31 No. 3, May 2007, 169–180
DOI: 10.1177/0146621606291569

& Stone, 1979). Both classical test theory and item response theory methods of estimating conditional standard errors of measurement have the mutual goal of providing a local measure of precision along the score scale. For reasons of space limitations, only the conceptual underpinnings of the classical test theory/strong true-score binomial error and item response theory models are highlighted here. Interested readers are encouraged to review the details of each method cited in the references above and also in Lord and Novick (1968) and Thissen and Wainer (2001).

The binomial error model is a score-based model where observed scores and true scores remain similarly defined as in the classic true-score model. In fact, according to Allen and Yen (1979), the two models are analogous in the following ways.

The binomial (or compound binomial, when locally independent items vary in level of difficulty) error models share the following assumptions with the classical true-score model: (a) The expected value (population mean) of the error scores for any examinee is zero, (b) the expected value of the product of error and true scores is zero, (c) observed score variance is the sum of true-score variance and error-score variance, (d) the squared correlation between observed and true scores is the ratio of true-score variance to observed-score variance, and (e) the squared correlation between observed and true scores is 1 minus the ratio of error-score variance to observed-score variance (i.e., when error variance is small, relative to observed-score variance, the proportion of variance in true scores explained by observed scores is large).

Two differences between the classical true-score model and the binomial error model are (a) the probability of an examinee obtaining a correct response is able to be estimated, and (b) in the binomial model, errors of measurement are allowed to vary by true-score level, thereby allowing for the estimation of conditional errors of measurement across the score scale. One final assumption is that the conditional distribution of an observed score for a given true score is the binomial or compound binomial distribution (Lord & Novick, 1968, p. 508). Estimates produced by the binomial error model provide similar coefficients of score reliability as those obtained in the classical true-score model. Indeed, a lower bound estimate of reliability, formally known as the Kuder-Richardson formula 21, is obtained when the regression of the true score on the observed score is observed to be linear (Wainer & Thissen, 2001, p. 61).

Item response theory (IRT) provides an alternative framework for modeling the association between an individual's response to an item and the underlying trait being measured. Specifically, the process of scaling data within the item response theory framework yields nonlinear regression curves that reflect the true relationship between examinees' observed responses and the latent variable that the test is purporting to measure. A particularly useful feature of the IRT modeling approach is that for every examinee, an estimate of ability, θ, is provided along with an associated standard error. The IRT-derived conditional errors of measurement, given an examinee's ability, provide the requisite information for examining how errors of measurement vary across the score scale. For a detailed explanation on item response theory, readers are referred to Lord (1980) and Hambleton and Swaminathan (1985).

Another concept that plays a very important role in classical test theory (CTT) is the overall estimate of reliability for a group of examinees on a particular test (ρ_{xx}), also known as "score reliability" (Thompson, 2003, p. 3). As Thompson (1994) notes, estimates of score reliability derived from a test are group dependent; that is, "the same measure, when administered to more homogeneous or heterogeneous groups of subjects, will yield scores with differing levels of reliability" (p. 839). Score reliability, as defined at the group level, is expressed as the ratio of true-score variance to observed-score variance for a group/population of examinees. Reliability coefficients based on scores obtained from the administration of a particular test are also defined as the square of the correlation between observed scores and true scores. According to this definition, score reliability or the reliability of measurement varies between 0 and 1, with reliability coefficients

ranging from between the lower .70s and into the high .90s as being desirable, depending on the type of instrumentation used and research setting within which the measurement occurred (Cohen & Swerdlik, 2005). Information about score reliability is routinely used in evaluating the psychometric quality of a test. Unfortunately, at the present time, there is no examinee-level score reliability or conditional reliability of scores obtained on a test. That is, there is only group-level reliability, and hence all examinees in a group are assumed to have the same score reliability, irrespective of their observed scores and true scores. Recently, a rationale for reporting estimates of conditional score reliability was provided in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999). Standard 2.1 claims the following: "For each total score, subscore, or combination of scores that is to be interpreted, estimates of reliabilities and standard errors of measurement or test information functions should be reported" (p. 31). In view of the usefulness of conditional SEM_ss noted above, it would be desirable to have a definition of conditional score reliability that is also on the same metric as the group-level reliability coefficient. The purpose of this investigation is to offer a definition of conditional reliability and illustrate it with empirical examples.

A Definition of Conditional Reliability

Prior to offering a definition of conditional reliability, the relationship between group-level SEM^2 and examinee-level (or conditional) SEM_s^2 is examined.

Group-Level SEM^2

According to Lord and Novick (1968), the group-level SEM^2 is the average (or expectation) of the examinee-level SEM_s^2s. That is,

$$SEM^2 = E(SEM_s^2), \tag{5}$$

where the expectation is taken over all examinees in a group. According to equation (5), some examinee-level SEM_s^2s may be smaller than the group-level SEM^2, and others may be larger.

Group-Level Reliability

The definition of group-level score reliability in the CTT framework may be stated as follows (Lord & Novick, 1968):

$$\rho_{xx} = \frac{\sigma_t^2}{\sigma_x^2} = \frac{\sigma_x^2 - SEM^2}{\sigma_x^2}, \tag{6}$$

where σ_x^2 and σ_t^2 represent the observed-score variance and true-score variance, respectively, in the group. The numerator on the extreme right of equation (6) is based on one of the well-known identities in CTT, which can be stated as

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 = \sigma_t^2 + SEM^2. \tag{7}$$

In view of equation (5), equation (6) can be rewritten as

$$\rho_{xx} = \frac{\sigma_x^2 - E(SEM_s^2)}{\sigma_x^2}. \tag{8}$$

Because σ_x^2 is the variance of observed scores of a group of examinees (e.g., the group-level variance on a standardized test), it does not vary from examinee to examinee. Therefore,

$$\rho_{xx} = \frac{\sigma_x^2 - E(SEM_s^2)}{\sigma_x^2} = E\left(\frac{\sigma_x^2 - SEM_s^2}{\sigma_x^2}\right). \tag{9}$$

The last part of this equation forms the basis for examinee-level score reliability.

Examinee-Level Reliability

It is proposed that the examinee-level reliability be defined as

$$\rho_{x_s x_s} = \frac{\sigma_x^2 - SEM_s^2}{\sigma_x^2}. \tag{10}$$

Several important features should be noted about this definition of conditional reliability. First, on the right-hand side of the above equation, for a particular test (e.g., a standardized test of achievement or ability, as in this investigation), only SEM_s^2 varies from one examinee to the next, whereas the observed-score variance remains fixed. Second, the conditional reliability will change as a function of the observed-score variance on a particular test. That is, when the observed-score variance of a test changes (e.g., when equation (10) is used with more or less heterogeneous data), the associated conditional reliability will also change accordingly. Therefore, clearly explaining and reporting the situation-specific observed-score variance and its role in the derivation of the examinee- or score-level conditional reliability is essential for proper interpretation of the index. Third, given a fixed group-level observed-score variance for a particular test, an examinee with a larger SEM_s^2 will have a smaller conditional reliability; conversely, a person with a smaller SEM_s^2 will have a larger examinee-level (or conditional) reliability. Finally, according to equations (9) and (10), the average (or expectation) of examinee-level (or conditional) reliability is equal to the group-level reliability. That is,

$$E(\rho_{x_s x_s}) = E\left(\frac{\sigma_x^2 - SEM_s^2}{\sigma_x^2}\right) = \frac{\sigma_x^2 - E(SEM_s^2)}{\sigma_x^2} = \rho_{xx}, \tag{11}$$

where the expectation is taken over all examinees in a group. Equations (5) and (11) enjoy a very important, common property: According to equation (5), the group-level SEM^2 is the average of the examinee-level SEM_s^2, whereas, according to equation (11), the group-level reliability is the average of examinee-level reliabilities. The examinee-level reliability will (mostly) vary between 0 and 1. When $SEM_s^2 = 0$ for examinee s, this examinee's reliability will be 1, as it should be. Because SEM_s^2 cannot be negative, examinee-level reliability can never be greater than 1. Examinee-level reliability will be greater than 0 except under some very special conditions. Consider a scenario in which all examinees have the same true score. Then, according to equation (7), the observed-score variance (σ_x^2) will be equal to SEM^2, and hence some of the individual SEM_s^2 may be greater than the average SEM^2, which, in turn, will result in a negative conditional reliability for some examinees, according to equation (10). Because the group-level reliability is zero when all examinees have the same true score (equation (6)), it may make sense to set all conditional reliabilities to zero in this very unlikely scenario. In most practical situations, however, the variance of true scores will be nonzero and substantial, and hence the conditional reliabilities will be greater than 0.

The conditional reliability defined in equation (10) does not conform to the usual definition of score reliability, which is the ratio of true-score variance to observed-score variance. For a given examinee, the true-score variance is zero because an examinee's true score is assumed to be a constant over replications. The examinee's observed scores are likely to vary over replications, thus resulting in a nonzero observed-score variance. Therefore, for a given examinee, the ratio of true-score variance to observed-score variance will be zero. The definition of examinee-level score reliability, given in equation (10), provides a solution to this problem and offers a practically useful measure of reliability for an examinee.

Finally, group-level reliability may be viewed as a standardized measure of group-level SEM^2. As noted earlier, SEM^2 or its square root is very useful in practice, even more useful than a measure of group-level reliability. Unfortunately, because of metric differences, it is difficult to compare group-level SEM^2s from two different tests; the metric of SEM^2 is dependent (within CTT) on the number-correct score metric of the test at hand. One way to make it possible for practitioners to compare SEM^2s across tests is to put them on a common metric. According to equation (7), SEM^2 can never be greater than σ_x^2 and, as a variance, can never be less than zero. Therefore, a standardized group-level SEM^2 may be defined as follows:

$$Standardized\ SEM^2 = \frac{(Maximum\ Possible\ SEM^2) - (SEM^2)}{(Maximum\ Possible\ SEM^2)} = \frac{\sigma_x^2 - SEM^2}{\sigma_x^2}, \qquad (12)$$

which is identical to the reliability measure given in equation (6) for a group. In view of this alternate definition of group-level reliability, examinee-level (or conditional) reliability given in equation (10) may be thought of as the standardized examinee-level SEM^2 using the same standardization process inherent in equation (12). As previously noted, the average of conditional reliabilities is equal to the group-level reliability, which is widely accepted in the psychometric community.

Conditional Reliability Within the IRT Context

The conditional reliability proposed above within the CTT framework is equally appropriate within the IRT context. The group-level reliability within IRT may be expressed as follows (Lord, 1983; Samejima, 1994):

$$\rho_{\hat{\theta}\hat{\theta}} = \frac{\sigma_{\theta}^2}{\sigma_{\hat{\theta}}^2} = \frac{\sigma_{\hat{\theta}}^2 - E(SEM_s^2)}{\sigma_{\hat{\theta}}^2}, \qquad (13)$$

where $\hat{\theta}$ is an estimate of ability (θ) for an examinee and, as before, the expectation (E) is taken over all examinees in a group. The variance of $\hat{\theta}_s$ ($\sigma_{\hat{\theta}}^2$) can be obtained by computing the variance of estimated abilities of examinees in a given group. The expectation of SEM_s^2 may be estimated by first computing the SEM_s^2 for each examinee (using his or her estimate of theta) and then taking the average of SEM_s^2s. Because the IRT-based SEM_s^2 is inversely (and asymptotically) equal to the test information function, the SEM^2 for examinee s can be obtained from

$$SEM_s^2 = \frac{1}{I_s}, \qquad (14)$$

where I_s is the total test information function for examinee s. The test information function is the sum of item information functions, and the formulas for computing the item and test information functions for the one-, two-, and three-parameter logistic IRT models may be found in Hambleton and Swaminathan (1985) and Lord (1980).

In view of equation (13), the examinee-level reliability within the IRT context may be expressed as

$$\rho_{\hat{\theta}_s\hat{\theta}_s} = \frac{\sigma^2_{\hat{\theta}} - SEM^2_s}{\sigma^2_{\hat{\theta}}}. \tag{15}$$

This equation in the IRT context is similar to equation (10) in the CTT context.

Two Illustrations

Having defined examinee-level reliabilities in the CTT and IRT frameworks, these definitions can now be illustrated with data from two tests, a picture completion test and a mathematics test.

Picture Completion Test

The picture completion test is a subtest of the Wechsler Preschool and Primary Scale of Intelligence–Third Edition (WPPSI-III; Wechsler, 2002). It is intended for children ages 4.0 to 7.3, and it is designed to measure a child's visual perception and organization, concentration, and visual recognition of essential details of objects. The subtest consists of 32 dichotomously scored items; for all items, children are required to view a picture and point to or name the important part missing from that picture. The sample for the current analysis consists of 1,100 examinees, drawn from the 2002 standardization of the WPPSI-III. The group-level number-correct score and IRT summary statistics are shown in Table 1. With respect to IRT, the data for the picture completion test were calibrated with the one-parameter (Rasch) model using PARSCALE, Version 4 (Muraki & Bock, 2002).

Conditional *SEM*s and reliabilities as well as the Rasch ability estimates for the picture completion test are shown in Table 2. The CTT-based conditional statistics are shown on the left and the IRT-based statistics on the right. The conditional *SEM*s within the CTT framework were derived based on the binomial error model (Lord & Novick, 1968), as implemented in the computer program BINSS (Brennan, 1997). It should be noted that there are several other methods for computing the CTT-based *SEM*s (Feldt et al., 1985; Qualls-Payne, 1992), but given the illustrative nature of this part of the investigation, only one method was used. The CTT- and IRT-based conditional reliabilities were computed using equations (10) and (15), respectively.

The conditional reliabilities in Table 2 range from .83 to .98 for the CTT framework and from .68 to .94 for the IRT framework. Although these ranges appear to be quite comparable, there is an important distinction. Within the CTT framework, high conditional reliabilities appear at the high end and the low end of the number-correct score scale, whereas low conditional reliabilities appear in the middle. The converse is true within the IRT framework; that is, low conditional reliabilities appear at the high and low ends of theta scale, with high conditional reliabilities appearing in the middle. This phenomenon is consistent with what is known about conditional *SEM*s in both frameworks. Conditional *SEM*s are typically lower in the middle for the IRT-based ability scale; therefore, the conditional reliabilities are higher in the middle. Similarly, conditional *SEM*s are lower at the extremes in the number-correct score metric; therefore, the corresponding conditional reliabilities are higher at the upper and lower ends of the number-correct score scale. This reflects the nonlinear relationship that exists between the number-correct score metric (in CTT) and the theta metric (in IRT). Additional information on nonlinear transformations and their effect on conditional *SEM*s can be found in Kolen, Hanson, and Brennan (1992).

As previously noted, the average of all conditional reliabilities, either in the CTT framework or in the IRT framework, is equal to the group-level reliability. As shown in Table 1, the group-level

Table 1

Number-Correct Score and IRT Summary Statistics at the Group Level

	Example 1[a]	Example 2[b]
Sample size	1,100	1,000
Number of items	32	40
Number-correct score		
Mean	18.49	19.6
SD	7.06	8.34
Variance[c]	49.87	69.55
Reliability[d]	.89	.89
SEM	2.34	2.77
IRT (theta)		
Mean	0.00	0.00
SD	1.00	0.95
Reliability[e]	.90	.85
SEM	0.31	0.37

Note. *SEM* = (group-level) standard error of measurement; IRT = item response theory.
a. Picture completion test (Wechsler Preschool and Primary Scale of Intelligence–Third Edition).
b. Mathematics test.
c. (Group-level) observed-score variance.
d. (Group-level) alpha reliability based on classical test theory (CTT).
e. (Group-level) reliability based on Formula 13.

CTT and IRT reliabilities are .89 and .90, respectively. It should be noted that the average means the average of the conditional reliabilities of all examinees in the sample, not just the average of the conditional reliabilities listed in columns 3 and 6 of Table 2. Finally, the Rasch ability estimates for the picture completion test are shown in column 5 of Table 2. Because the number-correct score is a sufficient statistic in the Rasch model, there is only one theta estimate associated with each possible number-correct score (Wright & Stone, 1979); that is, the correspondence between columns 1 and 5 is one-to-one.

Mathematics Test

This test consists of 40 multiple-choice, dichotomously scored items. It is part of the New Hampshire Educational Improvement and Assessment Program. The current data set consists of a sample of 1,000 examinees, drawn from the 2003 administration of the Grade 10 mathematics assessments. This portion of the assessment is a general mathematics program that measures probability and statistics, discrete math, numeration, operations, number theory, geometry, measurement, trigonometry, functions, and algebra. The assessment consists of several test forms, and only multiple-choice items that were common across all test forms were used in this study.

Group-level number-correct score and IRT summary statistics for the mathematics test are also shown in Table 1. The three-parameter logistic (3-PL) IRT model was used to calibrate this test with the help of the BILOG-MG3 computer program (Zimowski, Muraki, Mislevy, & Bock, 2002). Conditional *SEM*s and reliabilities for this test are shown in Table 3. As in the previous example, the conditional *SEM*s in the CTT framework were obtained with the binomial error model (Lord & Novick, 1968). IRT-based conditional *SEM*s were directly obtained from the

Table 2

Conditional *SEM*s and Reliabilities for the Picture Completion Test

CTT				IRT 1−PL (Rasch) Model			
Number-Correct Score	Conditional *SEM* (Binomial Error Model)	Conditional Reliability	Observed-Score Variance	Theta	Conditional *SEM*	Conditional Reliability	Observed-Score Variance
1	1.00	.98	49.87	−3.81	0.57	.68	49.87
2	1.39	.96	49.87	−2.94	0.48	.77	49.87
3	1.68	.94	49.87	−2.72	0.43	.82	49.87
4	1.90	.93	49.87	−2.39	0.39	.85	49.87
5	2.08	.91	49.87	−2.11	0.37	.86	49.87
6	2.24	.90	49.87	−1.83	0.34	.88	49.87
7	2.37	.89	49.87	−1.62	0.33	.89	49.87
8	2.48	.88	49.87	−1.52	0.32	.90	49.87
9	2.58	.87	49.87	−1.34	0.31	.90	49.87
10	2.66	.86	49.87	−1.25	0.30	.91	49.87
11	2.73	.85	49.87	−0.93	0.27	.93	49.87
12	2.78	.85	49.87	−0.79	0.26	.93	49.87
13	2.82	.84	49.87	−0.66	0.25	.94	49.87
14	2.85	.84	49.87	−0.64	0.26	.93	49.87
15	2.86	.84	49.87	−0.41	0.25	.94	49.87
16	2.87	.83	49.87	−0.29	0.24	.94	49.87
17	2.86	.84	49.87	−0.17	0.24	.94	49.87
18	2.85	.84	49.87	−0.05	0.24	.94	49.87
19	2.82	.84	49.87	0.07	0.24	.94	49.87
20	2.78	.85	49.87	0.19	0.25	.94	49.87
21	2.72	.85	49.87	0.31	0.25	.94	49.87
22	2.66	.86	49.87	0.44	0.25	.94	49.87
23	2.58	.87	49.87	0.57	0.26	.93	49.87
24	2.48	.88	49.87	0.70	0.26	.93	49.87
25	2.37	.89	49.87	0.84	0.27	.93	49.87
26	2.24	.90	49.87	0.99	0.27	.93	49.87
27	2.08	.91	49.87	1.15	0.29	.92	49.87
28	1.90	.93	49.87	1.34	0.30	.91	49.87
29	1.67	.94	49.87	1.54	0.34	.88	49.87
30	1.39	.96	49.87	1.81	0.39	.85	49.87
31	1.00	.98	49.87	2.22	0.41	.83	49.87
32	1.00	.98	49.87	—	—	—	

Note. *SEM* = standard error of measurement; CTT = classical test theory; IRT = item response theory; 1-PL = one-parameter logistic model. "—" indicates that theta and the associated standard error was unable to be estimated in the Rasch model due to a respondent obtaining a perfect score.

BILOG-MG3 output. As before, the CTT- and IRT-based conditional reliabilities were computed with the help of equations (10) and (15), respectively.

As shown in Table 3, the conditional reliabilities go from a low of .85 to a high of .99 in the CTT framework, and they vary from .63 to .99 in the IRT framework. As in the previous example, the higher conditional reliabilities are associated with extreme number-correct scores (high and

Table 3

Conditional *SEM*s and Reliabilities for the Mathematics Test

CTT				IRT 1−PL (Rasch) Model			
Number-Correct Score	Conditional *SEM* (Binomial Error Model)	Conditional Reliability	Observed-Score Variance	Theta	Conditional *SEM*	Conditional Reliability	Observed-Score Variance
1	1.00	.99	69.55	−1.97	0.57	.64	69.55
2	1.40	.97	69.55	−1.95	0.57	.63	69.55
3	1.69	.96	69.55	−1.62	0.50	.72	69.55
4	1.92	.95	69.55	−1.75	0.55	.67	69.55
5	2.12	.94	69.55	−1.62	0.53	.69	69.55
6	2.29	.92	69.55	−1.60	0.53	.69	69.55
7	2.43	.92	69.55	−1.50	0.53	.69	69.55
8	2.56	.91	69.55	−1.32	0.54	.67	69.55
9	2.67	.90	69.55	−1.22	0.53	.68	69.55
10	2.77	.89	69.55	−1.15	0.54	.68	69.55
11	2.86	.88	69.55	−1.00	0.50	.72	69.55
12	2.94	.88	69.55	−0.82	0.46	.77	69.55
13	3.00	.87	69.55	−0.73	0.42	.80	69.55
14	3.06	.87	69.55	−0.62	0.37	.85	69.55
15	3.10	.86	69.55	−0.59	0.39	.83	69.55
16	3.14	.86	69.55	−0.41	0.35	.86	69.55
17	3.17	.86	69.55	−0.22	0.39	.83	69.55
18	3.19	.85	69.55	−0.08	0.41	.81	69.55
19	3.20	.85	69.55	0.05	0.40	.82	69.55
20	3.20	.85	69.55	0.17	0.34	.87	69.55
21	3.20	.85	69.55	0.36	0.24	.94	69.55
22	3.19	.85	69.55	0.41	0.17	.97	69.55
23	3.17	.86	69.55	0.43	0.12	.98	69.55
24	3.14	.86	69.55	0.45	0.09	.99	69.55
25	3.10	.86	69.55	0.46	0.14	.98	69.55
26	3.05	.87	69.55	0.49	0.20	.96	69.55
27	3.00	.87	69.55	0.71	0.36	.86	69.55
28	2.94	.88	69.55	0.89	0.38	.84	69.55
29	2.86	.88	69.55	1.12	0.34	.87	69.55
30	2.77	.89	69.55	1.22	0.28	.91	69.55
31	2.67	.90	69.55	1.30	0.16	.97	69.55
32	2.56	.91	69.55	1.33	0.08	.99	69.55
33	2.43	.92	69.55	1.34	0.09	.99	69.55
34	2.29	.92	69.55	1.38	0.19	.96	69.55
35	2.12	.94	69.55	1.45	0.28	.91	69.55
36	1.92	.95	69.55	1.81	0.43	.80	69.55
37	1.69	.96	69.55	1.98	0.39	.83	69.55
38	1.40	.97	69.55	2.22	0.31	.89	69.55
39	1.20	.98	69.55	2.34	0.36	.85	69.55
40	1.00	.99	69.55	2.59	0.49	.73	69.55

Note. *SEM* = standard error of measurement; CTT = classical test theory; IRT = item response theory; 1-PL = one-parameter logistic model.

low) in the CTT framework, whereas the lower conditional reliabilities are in the middle of the number-correct score metric. Also, as previously noted, the converse is true for the IRT framework. There is, however, one important difference between the two examples. The mathematics test was calibrated with the 3-PL model and, hence the number-correct score is not a sufficient statistic. Therefore, the estimates in the theta column do not directly correspond to the number-correct scores in the first column. For example, in Table 3, a number-correct score of 4 is associated with a theta of –1.75. This theta estimate is the average of all theta estimates of examinees whose number-correct score is 4. The other theta estimates in Table 3 are to be similarly interpreted. Furthermore, conditional *SEM*s are similarly obtained; that is, the reported conditional *SEM* is the square root of the average of squared *SEM*s of all those examinees with the same number-correct score. All 1,000 theta estimates could have been reported (and please note that these theta estimates, in theory, could be different from one examinee to the next in the two-parameter logistic [2-PL] and 3-PL models), as well as the associated conditional *SEM*s and reliabilities, but that would have been too much data to report, so the average theta and the average conditional *SEM* for a given number-correct score were chosen. This was done only for illustrative purposes. Interested readers can obtain all possible theta estimates, conditional *SEM*s, and conditional reliabilities from one of the authors.

In Table 2, the number-correct score of 6 corresponds to a conditional reliability of .90 within the CTT framework. Similarly, in Table 3, a number-correct score of 9 also corresponds to a conditional reliability of .90. From a measurement precision point of view, these two number-correct scores (6 from the picture completion test and 9 from the mathematics test) are equally reliable. Such an inference is not so readily available when one is restricted to use only conditional *SEM*s because the conditional *SEM* is 2.24 for number-correct score 6 in the picture completion test and 2.67 for number-correct score 9 in the mathematics test. The conditional *SEM*, among other things, is a function of the number of items in the test and the number-correct score metric. The same is equally true in the IRT metric because, among other things, the conditional *SEM* depends on the number of items in the test.

Summary and Conclusions

It has been noted that the group-level SEM^2 is simply the average of all conditional SEM^2s and that the group-level test score reliability can be viewed as a standardized group-level SEM^2 so that it varies between 0 and 1. The variance of number-correct scores (in the CTT framework) or the variance of thetas (in the IRT framework) is used as the basis for standardizing the group-level SEM^2. The same standardization procedure, when applied to conditional SEM^2s, leads to a definition of conditional reliability. The conditional reliability, so defined, could vary from person to person, and the average of such conditional reliabilities would equal the group-level reliability in both CTT and IRT frameworks. Information about conditional reliability could be very useful for practitioners in evaluating the measurement precision of individuals within a test or across tests because conditional reliability, like group-level reliability, is on a common metric ranging from 0 to 1. Two examples are presented to illustrate the computation and the potential usefulness of conditional reliabilities.

This study has introduced the concept of conditional reliability and has described a method for computing reliability for each unique total number-correct score point or ability point on an assessment. With the increased amount of standardized assessments throughout educational and psychological research communities, conditional reliability may have important implications for reporting of scores and for test development. The recommendation and associated need to report

estimates of conditional standard errors of measurement and reliability is clearly evidenced in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). Therefore, the work presented herein meets the need to report reliability for each score point and will serve an important role in interpreting the results of an assessment.

In Standard 2.3, ''when test interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability data, including standard errors, should be provided for such differences'' (AERA, APA, & NCME, 1999, p. 32). During a test development process, content experts, who are not necessarily familiar with IRT (or the technical underpinnings), may find conditional reliability coefficients useful when addressing this standard. For example, conditional reliability may be particularly useful for an assessment program where there is a critical total score (e.g., where a pass/fail decision is made). In this context, the test developer can optimize the performance of an assessment at the critical score level with a coefficient that he or she is already familiar with, rather than having to turn to target information functions. Therefore, the notion of conditional reliability, grounded in CTT and IRT and guided by the *Standards*, may be useful for measurement practitioners. Similarly, the fact that the CTT- and IRT-conditional *SEM*s and score reliability display opposite patterns in their distributions of errors allows researchers or test developers to critically examine the question of precisely where along the score continuum a particular method (CTT or IRT) yields the most reliable and valid estimates of local precision in relation to the purpose of the proposed measurement of the test.

The results of this investigation show how conditional reliability might be useful in applied settings. Further research is needed to demonstrate how the concept of conditional reliability might be used in a variety of assessment programs and help researchers understand any limitations of this concept. Nonetheless, the approach discussed in this article may be useful not only for developing tests as discussed above but also for reporting assessment results.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Brennan, R. (1997). *BINSS and CBINSS: Conditional standard errors of measurement for scale scores using binomial and compound binomial assumptions*. Iowa City: University of Iowa.

Cohen, R. J., & Swerdlik, M. (2005). *Psychological testing and assessment* (6th ed.). Boston: McGraw-Hill.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: American Council on Education/Macmillan.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, *29*, 285-307.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*, 233-245.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Muraki, E., & Bock, D. (2002). *PARSCALE 4: IRT based test scoring and item analysis for graded items and rating scales*. Chicago: Scientific Software.

Price, L. R., Raju, N. S., Lurie, A., Wilkins, C., & Zhu, J. (2004, August). *Conditional standard errors of measurement for composite scores on the WPPSI-III*. Poster session presentation at the annual meeting of the American Psychological Association, Division 5 (Quantitative Methods), Honolulu, HI.

Price, L. R., Raju, N. S., Lurie, A., Wilkins, C., & Zhu, J. (2006). Conditional standard errors of measurement for composite scores on the WPPSI-III. *Psychological Reports*, *98*, 237-252.

Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement*, *29*, 213-225.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, *18*, 229-244.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837-847.

Thompson, B. (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

Wainer, H., & Thissen, D. (2001). True score theory: The traditional method. In H. Wainer & D. Thissen (Eds.), *Test scoring* (pp. 23-72). Mahwah, NJ: Lawrence Erlbaum.

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence–Third Edition*. San Antonio, TX: The Psychological Corporation.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). BILOG-MG [Computer software]. St. Paul, MN: Assessment Systems Corporation.

Author's Address

Address correspondence to Larry R. Price, PhD Program in Education, #325 ASB South, Texas State University–San Marcos, San Marcos, TX 78666-4616; e-mail: lprice@txstate.edu.