PERSONAL STANDARD ERRORS OF MEASURMENT

Personal Standard Errors of Measurement

David F. Lohman and Rashid S. Al-Mahrazi

The University of Iowa

Abstract

It has long been recognized that the standard error of measurement (SEM) varies across score levels. Although SEMs conditioned on score level better caution score interpretations than a common SEM for all test takers, examinees with the same total score may have quite different SEMs. In this paper, we propose a general method for estimating SEMs for individual examinees for tests that have been scaled using Item Response theory (IRT). We then show how two variants of this procedure can be used to caution the scores of examinees whose patterns of item and subtest scores differ from the patterns expected when a unidimensional IRT model fits the data. The two variants are compared to estimates of SEM derived from generalizability theory and from classical test theory. The procedures are then illustrated using standardization data from Form 5 of the Cognitive Abilities Test (CogAT; Thorndike & Hagen, 1993).

Background

It has long been recognized that the standard error of measurement (SEM) is not the same for all examinees. Raw score SEMs are larger for scores near the mean than for extreme scores. Scale score SEMs typically show the opposite profile, although the pattern depends on the relationship between raw scores and scale scores (Brennan & Lee, 1999; Lee, Brennan, & Kolen, 2000). If SEMs vary across score levels, then one should not be equally confident in all scores on a test. Further, if SEM varies, then one cannot use a simple rule for deciding whether differences among subtest scores on a test battery are sufficiently large to warrant interpretation.

Interest in standard errors of measurement that are conditioned on score level—conditional standard errors of measurement (CSEMs)—has grown because of recommendations that test publishers report them (Joint Committee on Standards, 1999) and because of advances in methods for estimating CSEMs both for raw scores and for the scale scores typically reported to examinees (Brennan & Lee, 1999; Feldt, 1984; Feldt, Steffen, & Gupta, 1985; Kolen, Zeng, & Hanson, 1996; Lee, Brennan, & Kolen, 2000; Lord, 1984; Woodruff, 1990).

Although SEMs conditioned on score level better caution score interpretations than a common SEM for all test takers, examinees with the same total correct score may have quite different SEMs. Indeed, in an IRT framework, conditional standard errors can be understood as the expected value of the error distribution at a given ability level. Some examinees show somewhat smaller errors; others show larger errors. However, as Jarjoura (1986) observed: “The question of...[estimation] of measurement error for a particular examinee has not been studied to the same degree as average measurement error” (p. 175). He added that it is intuitively recognized that “an examinee-level error should result in larger error variance for an examinee who does much guessing than for one who does not (given both have the same true score).” Others have noted that

examinees vary in consistency. Some are consistently inconsistent; others are more consistently consistent in their behavior (Birdie, 1969). Jarjoura (1986) developed an estimator for examinee-level error using the framework of generalizability theory. The goal of this paper is to develop estimators for examinee-level error using the framework of item response theory (IRT). Two variants of an IRT-based SEM will be established for individual examinees. One variant gives estimates that are similar to the Jarjoura examinee-level error variance. The other estimate compares favorably with an even earlier SEM suggested by Thorndike (1951). The procedures developed here are illustrated using data from Form 5 of the Cognitive Abilities Test (CogAT; Thorndike & Hagen, 1993).

Item-level PSEM

For the three-parameter logistic IRT model (3PL), the probability that examinees at a specific ability level, θ, answer item i correctly is defined as,

$$p_i = \text{Prob}(x_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}, \quad i = 1, 2, \ldots, n, \tag{1}$$

where x_i is the observed response to item i,
a_i is the discrimination parameter for item i,
b_i is the difficulty parameter for item i,
c_i is the guessing parameter for item i,
n is the total number of items administered to the person.

The 2PL model fixes the c_i parameters to zero for all items and the 1PL model further fixes the a_i parameters for all items. Only estimates of both item and examinee ability parameters are available in practice. Here it is assumed that the maximum likelihood method is used to estimate examinee's ability scores using item pattern scoring. The true score of examinees with an ability estimate of θ is here defined as $\tau = \sum_{i=1}^{n} p_i$. The conditional error variance for examinees at a specific true score is defined as the expected value of the error distribution at a given true score, τ,

which is equal to the variance of the distribution of the total-correct scores at a given ability level, θ (Lord, 1980, p. 46). That is,

$$\sigma^2_{e/\tau} \equiv \sigma^2_{e/\theta} = \sigma^2_{x/\theta} = E\left(x - \sum_{i=1}^{n} p_i \right)^2 = \sum_{i=1}^{n} p_i q_i \ , \tag{2}$$

where x is the observed total-correct score, and $q_i = 1 - p_i$.

However, there are other ways to interpret the IRT conditional error variance. One way is to consider the IRT conditional error variance as the error variability in estimating an examinee's scores on individual items. That is,

$$\sigma^2_{e/\tau} \equiv \sigma^2_{e/\theta} = \sum_{i=1}^{n} p_i q_i = \sum_{i=1}^{n} E\left(x_i - p_i\right)^2 = E\left[\sum_{i=1}^{n} \left(x_i - p_i\right)^2 \right] \tag{3}$$

Equation 3 says that the IRT error variance is the expected value of the discrepancy between observed and predicted performance over all possible item score patterns that can be demonstrated by an examinee at a specific ability level. Hence, the IRT error variance represents the estimated error variance for a typical examinee at a specific ability level. Indeed, although it is highly improbable, a high ability examinee could respond incorrectly to all test items (total-correct score of zero). In general, even though examinees at a specific ability level can theoretically demonstrate any item score pattern, they are more likely to demonstrate certain item score patterns than other score patterns (Brennan & Lee, 1999).

Moreover, two examinees who have the same estimated ability level or true score but have different item score patterns could demonstrate different measurement error. For example, a low ability examinee who responds correctly to difficult items has a score with larger error variability on that measurement occasion than another examinee at the same ability level who responds to items in a manner that conforms to the expectations of the measurement model. Therefore, on any particular measurement occasion, some examinees show somewhat smaller errors, whereas others

at the same ability level show much larger errors. However, Equations 2 and 3 imply that these two examinees have the same error variability because they have the same ability level. The magnitude of this estimated error depends on the particular scaling model and scoring rules that are used. Changing either the scaling model (e.g., from 2PL to 3PL) or the scoring rule (e.g., from number correct to item pattern scoring) can change the estimated IRT SEM (Yen & Candell , 1991). Regardless of the magnitude of this estimate, however, it will be the same for all examinees who have the same estimated ability level. Because the IRT SEM reports only the expected error at each theta or true score, it treats as equivalent error variablities that we know are not equal.

The locus of these differences in error variabilities for persons at the same ability level is shown in Equation 3. The term in brackets in Equation 3 differs from one item score pattern to another. Some item score patterns will have larger values of that term, whereas others will have smaller values. This term can be used as an estimator of examinee-level or personal error variance based on an examinee *a*'s observed item score pattern at a given test administration. This personal error variance is most appropriate when the focus of the measurement is over examinee's scores on individual items. We call this the item-level personal error variance because it is based on examinee *a*'s observed item scores,

$$\hat{\sigma}^2_{e_a / \theta\,(item)} = \sum_{i=1}^{n} (x_{ai} - p_i)^2 \,, \tag{4}$$

where x_{ai} is the observed item *i* score for an examinee, *a*. The square root of the error variance in Equation 4 is the estimate of the item-level personal SEM (item-level PSEM).

Subtest-level PSEM

The examinee-level error variance in Equation 4 was established for those cases in which an examinee's scores on individual items are the focus of measurement. However, there are times when the focus of measurement is on examinee's total scores across groups of items, i.e., subtests.

Examples of possible subtests are content and format categories. For example, ability tests such as the SAT or the GRE present blocks of items in a common format such as sentence completions or analogies. The scores of examinees who show markedly different performance on different item types are not as dependable as those examinees who show approximately similar performance on these different item types. For such examinees, we are interested in estimating the conditional error variability associated with the examinee's scores across various subtests on the test. The examinee-level error variance in Equation 4 captures this type of error variability only indirectly.

The IRT error variance in Equation 2 can also be interpreted as the error variability in estimating the examinee's scores on various subtests, *j*, of the test. That is,

$$\sigma^2_{e/\tau} \equiv \sigma^2_{e/\theta} = \sum_{i=1}^{n} p_i q_i = \sum_{j=1}^{J} \left(\sum_{i=1}^{n_j} p_i q_i \right) = \sum_{j=1}^{J} E\left(x_j - \sum_{i=1}^{n_j} p_i \right)^2 = E\left[\sum_{j=1}^{J} \left(x_j - \sum_{i=1}^{n_j} p_i \right)^2 \right] \quad (5)$$

where $x_j = \sum_{i=1}^{n_j} x_i$ is the observed total-correct score over the n_j items of the subtest *j*,

$\sum_{i=1}^{n_j} p_i$ is the expected score on subtest *j*, which is the sum of the probabilities of correct responses to the n_j items of the subtest *j*, and

J is the total number of subtests.

Equation 5 shows that the IRT error variance is the expected value of the error over all possible subtest score patterns that can be demonstrated by an examinee at a specific ability level. However, the IRT error variance in Equation 2 cannot reveal the differences among examinees at the same ability level or true score when they have different subtest score patterns. It is expected that the examinee is more likely to demonstrate certain subtest score patterns more than others, although examinees at a specific ability level can theoretically demonstrate any subtest score patterns.

The term in brackets in Equation 5 differs from one subtest score pattern to another. Some

subtest score patterns will have larger values of that term, whereas others will have smaller values. Hence, this term can be used as an estimator of examinee-level or personal error variance based on the examinee's observed subtest score pattern at a given test administration. Because this error variance uses an examinee *a*'s scores on various subtests of the test, it will be referred to as subtest-level personal error variance,

$$\hat{\sigma}^2_{e_a / \theta\ (subtest)} = \sum_{j=1}^{J} \left(x_{aj} - \sum_{i=1}^{n_j} p_i \right)^2 , \tag{6}$$

where x_{aj} is the observed subtest *j* score for an examinee, *a*. The square root of the error variance in Equation 6 is the estimate of the subtest-level personal SEM (subtest-level PSEM).

Interpretation of Personal Error Variance

Examinees with large item-level PSEMs need not have large subtest-level PSEMs. In fact an examinee could have an estimate of the subtest-level PSEM that is less than, equal to, or larger than the estimate of the item-level PSEM. For example, examinees can have total-correct scores on subtests that are close to the expectations of the model, whereas their item scores within each subtest do not correspond to the expectations of the model. Such examinees would show subtest-level PSEMs that are smaller than item-level PSEMs.

When one or both of the observed error estimates for an individual are significantly larger than the error expected under the assumption of random disturbances, then we can reject the hypothesis that all disturbances were random events. Rather, such individuals may have systematic errors in their test scores. Some individuals, on the other hand, will obtain error estimates that are much smaller than the average error. These are individuals who, on this testing occasion, exhibited behavior that was unusually consistent across items or subtests. A smaller than expected PSEM may even reflect a general consistency in behavior on the tasks in question

(Birdie, 1969).

Relationships with Estimates of Conditional SEM in Classical Test Theory

The two variants of personal standard error proposed here bear obvious similarities with Jarjoura's (1986) examinee-level SEM in generalizabiilty theory, and with the Lord (1955, 1965) and Thorndike (1951) SEMs in the context of classical test theory. Understanding the similarities and differences between these models illuminates the advantages and limitations of each. We first outline these estimates and then relate each to the corresponding personal standard error that we propose.

Jarjoura's SEM. Jarjoura (1986) proposed an examinee-level error variance conditioned on examinee's true score, $\tau = \mu_a$. Jarjoura's SEM adjusts for the mean difficulty of items (mean proportion correct) on a particular test form. Jarjoura's examinee-level error variance for mean adjusted scores is defined as follows:

$$\sigma^2_{Jarjoura} = \underset{i}{E}(x_{ai} - \mu_a - \mu_i + \mu)^2 \tag{7}$$

where $\mu_a \equiv \underset{i}{E}(x_{ai})$ is the mean (proportion-correct) score for examinee *a* over the universe of items,

$\mu_i \equiv \underset{a}{E}(x_{ai})$ is the population mean of all examinees taking item *i*, and

$\mu \equiv \underset{ai}{E}(x_{ai})$ is the mean over both the population of examinees and the universe of items.

The biased[1] estimator of the examinee-level error variance conditioned on estimated examinee's true score, $\hat{\tau} = \bar{x}_a$ (examinee's proportion-correct score) is,

$$\hat{\sigma}^2_{Jarjoura} = \sum_{i=1}^{n} (x_{ai} - \bar{x}_a - \bar{x}_i + \bar{\bar{x}})^2 \tag{8}$$

where $\bar{x}_a = \sum_i x_{ai} / n$ is the mean observed score for examinee *a* over test items and is the

analog of μ_a,

$\bar{x}_i = \sum_a x_{ai} / A$ is the mean of the observed scores for item *i* over *A* examinees taking the test and is the analog for μ_i,

$\bar{\bar{x}} = \sum_{ai} x_{ai} / nA$ is the mean of the observed scores over both the sample of examinees and the sample of test items and is the analog of μ.

Jarjoura's estimate of examinee-level error variance is known also as the conditional relative error variance in the framework of generalizability theory (Brennan, 2001). Brennan (1998) argued that this estimate of conditional relative error variance (Equation 8) is valid when the sample size of examinees administered the test form is large.

Jarjoura's examinee-level error variance adjusts the examinee's item scores for the estimated mean item difficulty on the test. This estimate is based on the responses of the sample of examinees who were administered the test. This adjustment for the item mean difficulty is the same for all examinees. Jarjoura (1986) recognized the need to make different adjustments to examinees' item scores for test difficulty that would vary along the ability score scale. However, this is not easily done in generalizability theory. In the IRT framework, the probability of a correct response, p_i, provides an estimate of item difficulty that varies along the ability score scale. The p_i in the IRT model can be interpreted as the probability of a correct response to item *i* for a group of examinees at the same ability level (Hambleton & Swaminathan, 1985). For dichotomously scored items, p_i gives the item *i* mean score (mean difficulty) over examinees at the same ability level as examinee *a*, and can be considered as an IRT analogue of μ_i. Because $\bar{p}_a = \sum_i p_i / n$ gives the examinee *a*'s proportion-correct true score (Hambleton & Swaminathan, 1985; Lord, 1980), $\bar{p}_a$ can be interpreted as an IRT analogue of μ_a. Also, $\bar{p}_a = \sum_i p_i / n$ can be interpreted as an IRT

analogue of μ, since $\mu \equiv \underset{i}{E}\,\mu_i \equiv \sum_i \mu_i / n$. If we substitute these terms in the right side of Equation 7, we get an estimate of an examinee-level error variance at an examinee level. This estimate is the same as the estimate of the proposed item-level personal error variance in Equation 4,

$$\sum_{i=1}^{n} (x_{ai} - \bar{p}_a - p_i + \bar{p}_a)^2 = \sum_{i=1}^{n} (x_{ai} - p_i)^2 = \hat{\sigma}^2_{e_a} / \theta\,(item) \cdot$$

It can be seen that the proposed estimate of the item-level personal error variance in the IRT context is analogous to the Jarjoura's estimate of examinee-level error variance in the generalizability theory. However, the IRT-based adjustment for test difficulty is believed to be more personalized for a given examinee than the adjustment for test difficulty estimated from the total sample of examinees of different ability levels who are administered the test form. However, it is expected that the average of both Jarjoura's error variance and the item-level personal error variance will be similar when the sample size of examinees administered the test form is large.

Lord's SEM. Lord (1965) suggested an estimate of conditional error variance for an examinee at a given true score estimated by the examinee's total-correct score, $\hat{\tau} = \bar{x}_a$,

$$\hat{\sigma}^2_{Lord} = n\bar{x}_a(1-\bar{x}_a)\,.$$

The relationship between the item-level personal error variance and Lord's error variance can be established from the relationship between the expected value of Jarjoura's error variance and Lord's error variance. Jarjoura (1986) demonstrated that

$$E\hat{\sigma}^2_{Jarjoura} = n\bar{x}_a(1-\bar{x}_a) - \sum_{i=1}^{n} (\bar{x}_i - \bar{\bar{x}})^2\,, \tag{9}$$

where $\sum_{i=1}^{n} (\bar{x}_i - \bar{\bar{x}})^2$ is the sum of squares of the deviation of item means (difficulties) from the grand mean. If we replace the terms in Equation 9 with their corresponding terms in the IRT framework at examinee level, we get

$$n\bar{p}_a(1-\bar{p}_a)-\sum_{i=1}^{n}(p_i-\bar{p}_a). \quad (10)$$

Lord (1980) demonstrated that Equation 10 is equal to the IRT error variance (Equation 2). Therefore, the expected value of Jarjoura's examinee-level error variance is equal to the expected value of the item-level personal error variance and, hence, equal to the IRT error variance. Moreover, this suggests that the IRT error variance is approximately equal to Lord's error variance adjusted for variation in item difficulties. The IRT error variance will be always smaller than Lord's error variance unless the test items have the same difficulty level.

Thorndike's SEM. Thorndike (1951) was among the first to propose a method for estimating conditional standard errors. He noted that error variance at different score levels could be estimated by taking the difference *D* between each examinee's scores on two parallel tests, x_{a1} and x_{a2}. Thorndike showed that the variance of the difference scores *D* was equal to twice the conditional error variance. Given a sufficiently large sample, one can estimate the conditional error variance at each level of x_a. More commonly, the observed total-correct score scale is divided into 8 to 10 intervals, and error variance is estimated for each score interval.

Thorndike's (1951) difference score procedure for estimating conditional error variance may be viewed as a special case of the more general method of computing a within-person error variance across *J* tau-equivalent test parts. Since test parts are assumed to be tau-equivalent, the expected score for each person is simply the mean across all test parts. The biased estimate of the Thorndike within-person error variance, here called $\hat{\sigma}^2_{Thorndike}$, is

$$\hat{\sigma}^2_{Thorndike}=\sum_{j=1}^{J}(x_{aj}-\frac{x_a}{J})^2=\sum_{j=1}^{J}(x_{aj}-\frac{n}{J}\bar{x}_a)^2. \quad (11)$$

However, Thorndike's within-person error variance is appropriate only if subtests have the same number of items. A modified expression can be suggested to estimate the within-person error

variance for subtests that are composed of a different number of items. The biased estimate of the modified within-person error variance, referred here as $\hat{\sigma}^2_{Mod.Thorndike}$, is

$$\hat{\sigma}^2_{Mod.Thorndike} = \sum_{j=1}^{J} (x_{aj} - n_j \frac{x_a}{n})^2 = \sum_{j=1}^{J} (x_{aj} - n_j \bar{x}_a)^2 . \quad (12)$$

If the subtests have the same number of items ($n_j = n/J$), then Equation 12 reduces to Equation 11.

The proposed subtest-level error variance has some similarities with Thorndike's within-person error variance. Whereas the expected score for each person in Thorndike's within-person error variance is the mean across all test parts, the expected score for each subtest in the proposed subtest-level error (Equation 6) is equal to the sum of the expected item scores for the items within the subtest. Deviations are computed about this expected score for each subtest rather than from the mean observed score across all test parts. The difference between the proposed subtest-level personal error variance and the modified expression of Thorndike's within-person error variance is that the proposed subtest-level personal error variance adjusts for examinee-level subtest difficulty, whereas the modified expression of the Thorndike within-person error variance does not. Therefore, the average subtest-level personal error variance is expected to be smaller than or equal to the modified expression of Thorndike's within-person error variance. These two estimates will be approximately equal if the subtests have the same mean difficulty.

The procedure Jarjoura (1986) used to obtain an examinee-level relative error variance can also be applied to the modified within-person error variance to adjust for subtest difficulty. Following the same process, the biased estimate of the relative modified within-person error variance is

$$\hat{\sigma}^2_{REL.Mod.Thorndike} = \sum_{j=1}^{J} (x_{aj} - n_j \bar{x}_a - n_j \bar{x}_j + n_j \bar{\bar{x}})^2 , \quad (13)$$

where $\bar{x}_j = \sum_{i=1}^{n_j} \bar{x}_i / n_j$ is the average score (measure of difficulty) for subtest *j*, which is equal to the average of an item's mean difficulty scores within the subtest *j*, and $\bar{\bar{x}}$ is the test mean score, which is equal to $\bar{\bar{x}} = \sum_{j=1}^{J} (n_j / n)\, \bar{x}_j$.

The relative modified within-person error variance adjusts the subtest error variability for the differences in the mean difficulty of the subtests based on a sample of examinees of different abilities administered the test form. Similar to Jarjoura's error variance (Equation 8), the estimate of the relative modified within-person error variance is valid when a large sample of examinees is administered the test form.

If the IRT terms ($\bar{p}_a$, p_i, $\bar{p}_a$) that are analogues to the three terms in Equation 8 ($\bar{x}_a$, $\bar{x}_i$, $\bar{\bar{x}}$, respectively) are substituted in Equation 13, we get an estimate of the relative modified within-person error variance adjusted for subtest difficulty at an examinee-level in the IRT context.

$$\sum_{j=1}^{J} (x_{aj} - n_j \bar{p}_a - n_j (\sum_{i=1}^{n_j} p_i) / n_j + n_j \bar{p}_a)^2 = \sum_{j=1}^{J} (x_{aj} - \sum_{i=1}^{n_j} p_i)^2 = \hat{\sigma}^2_{e_a} / \theta \ (subtest)$$

This estimate is the same as the estimate of the proposed subtest-level personal error variance in Equation 6. Hence, the proposed subtest-level personal error variance is analogous to the modified Thorndike within-person error variance after adjusting for subtest difficulty at an examinee-level. Similar to the item-level personal error variance, the subtest-level personal error variance is believed to be more personalized for a given examinee than the relative modified within-person error variance. However, it is expected that the average of both the relative modified within-person error variance and the subtest-level personal error variance will give similar estimates of examinee's error variability across subtest scores when the sample size of examinees administered the test form is large.

Applications

Each of the estimates of conditional SEM reported here was computed for 12,242 students who were administered the CogAT Level A Verbal battery as part of test standardization. This level of the test was chosen because it is typically administered at grade 3. This is the lowest level of the test at which students must read items, pace themselves, and record their responses on a separate machine-readable answer sheet. One might reasonably expect more children at this age to show confusion about how to solve particular subtests or how to keep track of their place on the answer sheet. This level of the test thus offers a good test case for the utility of personal standard errors. The verbal score that is reported for each child is based on the sum of the scores across three subtests: 20 items appear in a verbal classification format, 20 items in a sentence completion format, and 25 items in a verbal analogy format. Descriptive statistics on item and subtest difficulty for Level A of the CogAT Verbal Battery are presented in Table 1.

The estimations of item-level PSEM (square root of Equation 4) and subtest-level PSEM (square root of Equation 6) presume that the test has been scaled using any IRT model. Thorndike and Hagen (1992) used the Rasch model to scale the CogAT. We used this scaling, but also rescaled the test using 2PL and 3PL models for some analyses. Thetas for these models were estimated from examinees' item score patterns using maximum likelihood.

Results for the Item-level PSEM and Subtest-level PSEM

Descriptive statistics for the item-level PSEMs, the subtest-level PSEMs, and the IRT SEM are presented in Table 2 for the 1PL, 2PL, and 3PL IRT models. The average conditional SEM for each model in this table is the square root of the average conditional error variance. Although the average estimated SEM was similar for all three IRT models, item-level and IRT SEMs showed the expected decline from the 1PL to 3PL models (see, e.g., Yen & Candell, 1991).

Average subtest-level PSEMs, on the other hand, showed a different pattern. They were smallest for the 1PL model, but equal for two and 3PL models. More importantly, the average of the item-level PSEM was approximately the same as the average of the IRT SEM, suggesting that the item-level PSEMs and the IRT SEM reflect similar sources of error. However, the average of the subtest-level PSEM was substantially larger than the average of the IRT SEM, suggesting that cumulating item scores in this way provides a different perspective on error variability in the test battery. These hypotheses are examined in greater detail below.

The top panel in Figure 1 shows the plots of item-level PSEMs and the IRT SEM for all examinees at each true score using 1PL, 2PL, and 3PL models. The bottom panel in Figure 1 presents the corresponding plots for the subtest-level PSEMs and the IRT SEM. (Note that scale of the vertical axis differs between the top and bottom plots. The lines representing the IRT SEM in the two plots would be identical if the scales of the vertical axes in the two plots were equal). Both the item-level PSEM and subtest-level PSEM showed more scatter at the middle of the true score scale than at the two ends of the scale. The subtest-level PSEM showed more scatter than the item-level PSEM at all true scores. For example, for the 2PL model, the values of the subtest-level PSEM ranged from 0.117 to 16.596, whereas the values of the item-level PSEM ranged from 0.236 to 4.422 (see Table 2).

The item-level PSEMs showed the typical inverted U shape in which SEMs are high in the middle of the total-correct score scale and low at both ends of the scale. On the other hand, subtest-level PSEMs did not show an inverted U shape. The minimum values of the subtest-level PSEM at each true score were about the same and close to zero over the entire score range. Unlike the item-level PSEM, it is possible that an examinee can have a subtest-level PSEM close to zero regardless of the true score of the examinee. This can happen when the examinee's subtest scores

are close to the expectation of the subtest scores based on the model. Moreover, although the majority of the subtest-level PSEMs were moderate, there some were extremely high, especially in the middle of the score scale.

For the 3PL model, the IRT SEM was the same for the subset of examinees who had true scores equal to sum of the guessing parameters for all test items regardless their total-correct score or which item response patterns they demonstrated. However, the item-level PSEM gave different estimates for those examinees depending on their item score patterns. For this data set, these estimates were often smaller than the corresponding IRT SEM. However, in principle, the item-level PSEM could give estimates of error that are larger than or equal to the corresponding IRT SEM for the 3PL model. Similarly, the subtest-level PSEM also gave estimates of error that differed from the IRT SEM for this group of examinees under the 3PL model. For this data set, some of these subtest-level PSEM estimates were smaller than the corresponding IRT SEMs whereas other estimates were larger.

Correlations among estimates of conditional SEM for each of the three IRT models are presented in Table 3. These correlations confirmed previous results that showed only small differences among the three IRT models with respect to the patterns of both item-level and subtest-level PSEMs. Correlations for the item-level PSEMs were lowest for the 1PL-3PL comparison (r = 0.991) and highest for the 2PL-3PL comparison (r = 0.998). Although there were differences for some low-scoring examinees, it was clear that the item-level PSEM was least affected by the scaling model. Correlations among the three estimates of subtest-level PSEM were high, but not as high as for the item-level PSEM estimates. Once again, the lowest correlation was between PSEMs for the 1PL and 3PL models (r = 0.955), and the highest correlation was between PSEMs for the 2PL and 3PL models (r = 0.980).

Table 3 also presents the correlations among item-level PSEMs, subtest-level PSEMs, and IRT SEMs within each IRT model. The correlations between the item-level PSEMs and IRT SEM were high (0.932, 0.945, and 0.922 for 1PL, 2PL, and 3PL). However, the correlations of the subtest-level PSEM with the IRT SEM were low (0.391, 0.411, and 0.411 for 1PL, 2PL, and 3PL). The subtest-level PSEM also showed low correlations with the item-level SEM (0.429, 0.441, and 0.433 for 1PL, 2PL, and 3PL). Examination of the scatterplots of the item-level and subtest-level PSEMs for each model showed that violation of the assumption of linear relationship was not the cause of these low correlations. The scatterplots showed that while the subtest-level PSEMs were less scattered at low values of item-level PSEM, they were more scattered for large values of item-level PSEM. These patterns are expected given the relationships between SEMs and true scores shown in Figure 1. Figure 1 shows that the subtest-level PSEMs were more scattered than the item-level PSEMs at the middle of the true score scale where both item-level and subtest-level PSEMs were large. However, the subtest-level PSEMs were scattered approximately the same as the item-level PSEMs at both low and high true scores (where item-level and subtest-level PSEMs were small). These results indicate that the subtest-level PSEM and item-level PSEM give estimates of different types of measurement error even though their expected values are the IRT SEM.

Previous results showed that the subtest-level PSEMs were more scattered than the item-level PSEMs. Moreover, the maximum values for the subtest-level PSEM (16.669 for 1PL) were much larger than the maximum values of the item-level PSEM (4.486 for 1PL). But does this imply that the subtest-level PSEM is generally larger than the item-level PSEM? Comparing the values of item-level PSEM and subtest-level PSEM for all individuals reveals that the item-level PSEM was actually greater than the subtest PSEM for 57.0, 54.3, and 54.6 percent of the cases in

the 1PL, 2PL, and 3PL models, respectively. In fact, in analyses on other data sets (not reported here), we have found that the item-level PSEM is greater than the subtest-level PSEM to the extent that the scaling model does not fit the data. Clearly, the subtest-level PSEM is as frequently smaller than the item-level PSEM as it is larger. On the other hand, the subtest-level procedure is believed to give a much more realistic estimate of the SEM for those cases in which subtest scores differ markedly from each other. For these examinees, the several subtests are clearly giving quite disparate estimates of the examinees' ability. It would seem wise to caution these estimates with a larger SEM.

Comparison with Estimates of Conditional SEM in Classical Test Theory

Table 2 also presents summary statistics for the five estimates of conditional SEM that do not use IRT models. Again, the average conditional SEM is simply the square root of the average conditional error variances. Results showed that the average of the conditional relative SEM (Jarjoura's SEM = 3.256) was approximately equal to the average of the item-level PSEM. The average of Lord's SEM (3.497) was larger than the average of the conditional relative SEM (3.256) and the three item-level PSEMs by an amount related to the differences among item difficulty ($3.497 \approx [(3.256)^2+1.6268]^{1/2}$), where 1.6268 is the sum of squares of the deviation of item difficulties from the grand mean for the 65 items on the test. On the other hand, the average of the relative modified Thorndike within-person SEM (square root of Equation 13, average = 3.791) was approximately equal to the average of the subtest-level PSEM. However, the average of the modified Thorndike within-person SEM (square root of Equation 12, average = 3.942) was larger than the relative modified Thorndike within-person SEM and the average of the subtest-level PSEM because differences in average difficulty among the three subtests are considered error ($3.942 \approx [(3.791)^2+1.1703]^{1/2}$). The average of the Thorndike within-person SEM was large

(4.246) because of its inherent bias when subtests have different numbers of items.

Figures 2 and 3 show the average of all estimates of SEM at each total-correct score, separately for each method of estimating SEM. The IRT-based error estimates were also averaged at each total-correct score to facilitate the comparison with other error estimates in the classical test theory. Each of the three plots in Figure 2 presents the averages of all item-level SEMs, whereas each of the three plots in Figure 3 presents the averages of all subtest-level SEMs plus the IRT SEM. These plots are presented separately for the three IRT models. The three plots Figure 2 reveal that the Lord binomial SEM was higher than all other estimates at most total-correct scores. This is because the Lord estimate assumes that the probability of a correct response is the same for all items. Allowing the probability of a correct response to vary across items in the item-level PSEM and the IRT SEM resulted in lower standard errors. Adjusting for differences among item difficulties in the conditional relative SEM also resulted in lower standard errors. The averages of the item-level PSEMs and the conditional relative SEM showed a smoothed inverted U shape and were approximately the same as the averages of the IRT SEM. These three estimates of SEM coincided over most of the total-correct score scale for the 1PL and 2PL models. However, for the 3PL model, the IRT SEM had higher values for individuals with low total-correct scores. For all three models, the item-level PSEM had smaller values than the conditional relative SEM at all total-correct scores. This was most apparent at both low and high total-correct scores. Unlike Lord's SEM, the values of item-level PSEM and conditional relative SEM were larger than zero at the perfect score (0.077, 0.236, 0.130 for item-level PSEM with the 1PL, 2PL, and 3PL models, respectively, and 1.275 for the conditional relative SEM). Although no examinee obtained a zero score on the CogAT, it is expected that the value of the item-level PSEM would also be larger than zero for an examinee with zero total-correct score.

Table 3 shows that the correlations among the four estimates of item-level SEM were high. The pattern of the correlations among these four estimates of SEM reveals that the item-level PSEM correlated higher with the conditional relative SEM (r = 0.992, 0.989, 0.986, for 1PL, 2PL, 3PL), whereas the IRT SEM correlated higher with Lord's SEM (r = 0.998, 0.991, 0.970, for 1PL, 2PL, 3PL). These correlations were higher for the 1PL model because of the one-to-one correspondence between the total-correct scores and ability values. The 2PL and 3PL models do not affect these correlations substantially, however. This pattern of correlations supports the correspondence between the item-level PSEM and the conditional relative SEM (Jarjoura's SEM) and between the IRT SEM and Lord's SEM (or more accurately between the IRT SEM and Lord's SEM adjusted for the item difficulty).

Plots for the subtest-level SEMs are shown in the three panels on the right side of Figure 2. These plots show that the subtest-level PSEM and the relative modified Thorndike within-person SEM coincided at all total-correct scores for the 1PL, 2PL, and 3PL models. However, the subtest-level PSEM for the 3PL model had higher values for individuals with low total-correct scores because of the c_i parameter. Comparison of the plots for the item-level PSEM and subtest-level PSEM reveals that the subtest-level PSEM—unlike the item-level PSEM—was affected by the c_i parameter. The modified Thorndike within-person SEM, where the differences in average difficulty among the three subtests are not removed and are considered error, was higher than the relative modified Thorndike within-person SEM in all three plots. This bias in the modified Thorndike within-person SEM was not substantial, however, because of the small differences among the mean subtest difficulties (see Table 1).

The plots in Figure 2 also demonstrate the flaw in the Thorndike within-person SEM when subtests have different numbers of items (as the case with verbal CogAT data). The plots reveal

that the Thorndike within-person SEM had high values for individuals with high total-correct scores. As shown in the three plots, the modified Thorndike within-person SEM remedies this flaw. The averages of the four estimates of the subtest-level SEMs showed a bumpy inverted U shape. Moreover, all estimates of subtest-level SEM in the plots were higher than the IRT SEM for individuals with both low and medium total-correct scores. Again, these differences between the average of the subtest-level PSEM and the average of IRT SEM are not substantial given the scale of values of the subtest-level PSEM. However, this result could suggest also that there is a possibility of inconsistency between examinees' subtest observed scores on CogAT and the expected subtest scores based on both the IRT models and the observed subtest means.

With the exception of the Thorndike within-person SEM, the correlations among the four estimates of the subtest-level SEMs were high. The lower correlations of the Thorndike within-person SEM with other estimates were caused by the bias in this estimate when the numbers of items in the subtests are different. The relative modified Thorndike within-person SEM correlated (0.990, 0.986, 0.957 for 1PL, 2PL, 3PL) higher with the subtest-level PSEM than did the modified Thorndike within-person SEM (0.920, 0.913, 0.891 for 1PL, 2PL, 3PL). These correlations were higher with the 1PL models. However, the 2PL and 3PL scaling models reduced these correlations slightly. These correlations support the correspondence between the subtest-level PSEM and the relative modified Thorndike within-person SEM.

The correlations among the estimates of item-level SEM (item-level PSEM, IRT SEM, Lord's SEM, and conditional relative SEM) and the estimates of subtest-level SEM (subtest-level PSEM, modified Thorndike within-person SEM, relative modified Thorndike within-person SEM) presented in Table 3 were small (range between 0.379 and 0.441). These correlations suggest that the estimates of subtest-level SEM behave differently from the estimates of item-level

SEM. The item-level PSEM gives an estimator of item-based error that corresponds with the conditional relative SEM, whereas the subtest-level PSEM gives an estimator of error that corresponds most closely with the relative modified Thorndike within-person SEM.

Discussion

Psychologists have long been intrigued by the extent to which individuals differ in the consistency of their behavior. Some individuals are consistently inconsistent on repetitions of a task, whereas others show greater consistency in their performance (e.g., Berdie, 1969). Indeed, in their derivation of classical test theory, Lord and Novick (1968) explicitly allow for the possibility that "some persons' responses are inherently more consistent than those of others, and that we are able to measure some persons' responses more accurately than others" (p. 32). Nevertheless, errors of measurement are often assumed to be the same for all examinees or, more defensibly, for all examinees who obtain the same total-correct score. In this paper, two estimates of error of measurement for individual examinees are developed and illustrated. We call them personal standard errors of measurement (PSEMs) to distinguish them from more familiar conditional standard errors of measurement (CSEMs) or a common standard error of measurement (SEM). The first PSEM captures discrepancies between the observed pattern of item scores and the pattern predicted by a unidimensional IRT model. The second PSEM captures discrepancies between observed subtest scores and the subtest scores predicted by the IRT model. The subtest-level PSEM is particularly useful in detecting patterns of subtest scores that (for whatever reason) differ markedly from the patterns predicted by the scaling model. PSEMs vary over a wider range when estimated from subtest scores than when estimated from item scores. Importantly, these subtest-level PSEMs capture extreme variation in subtest scores. Operational testing programs may want to use one or both PSEM estimates. Test users who are interested in both estimates but

prefer to use one score that captures the advantages of both estimates could average the two estimates, or more conservatively, use the largest estimate for constructing confidence intervals around the total-correct scores for individual examinees.

The item-level PSEM behaved in the same way as the examinee-level SEM developed by Jarjoura (1986) or the conditional relative SEM (Brennan, 2001). However, the item-level PSEM proposed here generally results in smaller estimates of examinees' error variability. In the data set we examined here, 93 percent of individuals had values of the item-level PSEM (with the 1PL model) that were smaller than their values for the conditional relative SEM. This is probably because the item-level PSEM is more personalized for a specific individual examinee (assuming the data fit the selected IRT model). The item-level PSEM adjusts for item difficulties at the examinee level by computing the p_i for each test item only for examinees at the same ability level as the target examinee. However, the conditional relative SEM adjusts for the item difficulties that are estimated from all examinees in the sample. It uses only one estimate of item difficulty for all examinees.

In addition, there is a possibility of invalidly high values of conditional relative SEM caused by the nature of the estimate by itself. For example, if an examinee with $\bar{x}_a = 0.05$ guesses correctly the answer to an item with $\bar{x}_i = 0.342$ while $\bar{\bar{x}} = 0.636$ for the test form, then the value of the term in the conditional relative SEM for this item is

$$(1 - 0.05 - 0.342 + 0.636)^2 = (1 + 0.244)^2 = (1.244)^2 = 1.5475.$$

This value is outside the range of valid values for a term that ranges between 0 and 1. Only one such item is sufficient to cause the conditional relative SEM to overestimate the examinee-level SEM. On the other hand, this issue cannot arise with the item-level PSEM because values for p_i range between 0 and 1 regardless of how difficult items are or how able the examinee is. This

could explain the result presented in Figure 2 where the conditional relative SEM had a higher average than the item-level PSEM at all total-correct scores, especially for low and high total-correct scores.

Similarly, the subtest-level PSEM behaved in the same way as the relative modified Thorndike within-person SEM (Equation 13). Again, it is expected (using the same arguments as before) that more individuals would have values of subtest-level PSEM smaller than those of the relative modified Thorndike within-person SEM. However, in the data set used here, there were only 53.5, 45.5, and 47.9 percent of such individuals for 1PL, 2PL, and 3PL models. This could be explained by the small differences among the mean difficulties of the three subtests. Inspection of Equation 13 shows that variation in subtest difficulties (in Equation 13) would increase the size of the SEM.[2]

Of course, the major limitation of the subtest-level method is that each PSEM is typically estimated from a small number of subtest scores. With more subtests and more items in each, one gets a more stable (and generally smaller) estimate of error. Thus, one important topic for future research is how best to effect partitioning of an item string into test parts, given different assumptions about the source and magnitude of errors. But it does not follow that a more stable or smaller estimate of error is a better estimate of error. For any fixed collection of items that has been scaled, the vector of item scores for each examinee has an observed deviation from the response probabilities predicted by the model. If the scores on the test are generally well fit by the scaling model, then these deviations will be small and normally distributed. However, when systematic errors attach to particular items for some examinees and not others, then the error distributions will be skewed—especially if the partitioning of items into subtests sorts them into categories that capture these systematic errors. Without such partitioning, systematic errors are scattered

throughout the item set and treated as random error. But whether the deviations from the IRT model can be modeled as random events or are ascribed to a more systematic source, they are a fact of life for a particular examinee's responses on a given test. Further, as was shown for the CogAT data, if the subtests are of at least moderate length and all do in fact measure the same construct, then the subtest-level PSEMs can be quite useful in detecting deviations from the scaling model.

Subtest-level PSEM's can also be quite useful in signaling variation among subscores on achievement tests that report performance on clusters of items defined by particular objectives, or among subtests that are combined to create larger composites. For example, Language Usage composite scores are commonly defined by performance on spelling, capitalization, punctuation, and usage/expression subtests. Computation of a subtest-level PSEM would allow construction of a confidence interval around the composite Language Usage score that would caution interpretation for examinees who display unusual patterns of scores across subtests that are combined to form the composite. Such composite scores are surely less generalizable than the composites for individuals who behave more consistently across subtests.

Further, both item-level and subtest-level PSEMs can also be interpreted in the context of person fit analysis. Both estimates of PSEM quantify the amount of aberrance in examinee's responses from the expectations of the IRT model and thus can be used to detect person misfit. For example, a large increment to the item-level PSEM is obtained every time the examinee responds to an item *i* differently from the prediction of the IRT model that is quantified by p_i. Since the IRT SEM is the expected value of both the item-level PSEM and subtest-level PSEM, the unusualness of a given PSEM can easily be determined by comparing it to the IRT SEM at that score level. By obtaining the variance of the square of item-level PSEM (personal error variance), Al-Mahrazi (2003) developed and investigated two versions of a person fit index that compares the estimate of

item-level personal error variance to the IRT error variance. Al-Mahrazi (2003) found that this new person fit index that employs the item-level personal error variance performed better than Wright's (1977) mean square statistics and performed similar to or even better than the standardized likelihood index of Drasgow, Levine, & Williams (1985).

Nevertheless, there is an important difference between procedures described in this paper for estimating personal SEMs—and thus enabling the construction of confidence intervals around reported scores—and procedures for detecting person misfit in IRT models. Operational testing programs typically do not have the luxury of not reporting scores for examinees whose responses do not fit the scaling model. Drawing a bright line between fit and misfit seems to distort the continuous variation in misfit that is typically observed. Indices of model fit and confidence intervals also have different implications for score interpretation. Confidence intervals allow the user to judge how much scores are likely to vary on retest; model fit indices do not inform such judgments. Confidence intervals also allow more direct inferences about the reliability of observed differences in score profiles. Indeed, it was the search for better ways to caution users about large but unreliable differences among scores on the three CogAT batteries that lead to these methods.

References

Al-Mahrazi, R. S. (2003). Investigating a new modification of the residual-based person index and its relationship with other indices in dichotomous item response theory. Unpublished Doctoral Dissertation. The University of Iowa.

Berdie, R. F. (1969). Consistency and generalizability of intraindividual variability. *Journal of Applied Psychology, 53,* 35-41.

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement, 22,* 307-331.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L., & Lee, W. -C. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement, 59,* 5-24.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement, 44*, 883-891.

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

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

Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that

considers test form difficulty adjustment. *Applied Psychological Measurement, 10,* 175-186.

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

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.

Lee, W. -C., Brennan, R. L., & Kolen, M J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37,* 1-28.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement, 15,* 325-336.

Lord, F. M. (1965). A strong true score theory, with applications. *Psychometrika, 30,* 239-270.

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

Lord, F. M. (1984). Standard errors of measurement at different score levels. *Journal of Educational Measurement, 21,* 239-243.

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

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560-620). Washington, DC: American Council on Education.

Thorndike, R. L. (1982). *Applied psychometrics.* Itasca, IL: Riverside.

Thorndike, R. L., & Hagen, E. (1974). *Cognitive Abilities Test technical manual.* Boston: Houghton Mifflin.

Thorndike, R. L., & Hagen, E. (1993). *Cognitive Abilities Test (Form 5).* Chicago: Riverside.

Woodruff, D. J. (1990). Conditional standard error of measurement in prediction. *Journal of Educational Measurement, 25,* 191-208.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97-115.

Yen, W. M. & Candell, G. L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education, 4,* 209-228.

Footnotes

[1]Biased estimators are used with all estimates of SEM in the classical test theory framework in order to allow a direct comparison with estimates of SEM in IRT framework.

[2]Similarly, variation in item difficulties (in Equation 7) increases the size of the conditional relative error variance over the corresponding IRT-based estimate given in Equation 4.

Table 1

Summary descriptive of the difficulties of items and subtests on the CogAT Vocabulary Test.

	Min	Max	Average	Variance	Sum of Squares
Item-Level[a]	0.342	0.932	0.6363	0.0254	1.6268
Subtest-Level[b]	0.605	0.674	0.6363	0.0012 (0.5851)	0.0024 (1.1703)

[a] $n = 65$.
[b] 3 subtests: $n_1 = 20$, difficulty = 0.6382; $n_2 = 20$, difficulty = 0.6736; $n_3 = 25$, difficulty = 0.6050. Values in the parentheses are obtained using the scales of subtests' total-correct scores.

Table 2

Summary Results of Various Estimates of Conditional Standard Error of Measurement

Estimate and Model[a]	Min	Max	Average
Item			
1PL	0.077	4.486	3.222
2PL	0.236	4.422	3.202
3PL	0.130	4.469	3.193
Subtest			
1PL	0.102	16.669	3.762
2PL	0.117	16.596	3.786
3PL	0.054	16.392	3.786
IRT			
1PL	0.685	3.657	3.230
2PL	0.799	3.676	3.220
3PL	0.693	3.632	3.207
SEM estimates from classical test theory			
Rel	1.275	4.329	3.256
Lord	0.000	4.031	3.497
Thornd	0.000	17.146	4.246
Mod Thornd	0.000	16.690	3.942
Rel Mod Thornd	0.220	16.793	3.791

Notes. Item = item-level personal standard error of measurement (PSEM); Subtest = Subtest-level PSEM; CTT = classical test theory; IRT = item response theory SEM; Rel = Jarjoura's SEM; Thornd = Thorndike's SEM; Mod Thornd = modified Thorndike's within-person SEM; Rel Mod Thornd = relative modified Thorndike's within-person SEM.

[a] $n = 12{,}242$

Table 3

Intercorrelations among Various Estimates of Conditional Standard Error of Measurement

		Item			IRT			Subtest						
		1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL				
Item														
	1PL													
	2PL	0.994												
	3PL	0.991	0.998											
IRT														
	1PL	0.932												
	2PL		0.945		0.989									
	3PL			0.922	0.967	0.984								
Subtest														
	1PL	0.429			0.391									
	2PL		0.441			0.411		0.985						
	3PL			0.433			0.411	0.968	0.983					
Rel		0.992	0.989	0.986	0.934	0.941	0.917	0.430	0.431	0.422				
Lord		0.934	0.937	0.937	0.998	0.991	0.970	0.396	0.403	0.394	0.940			
Thornd		0.132	0.139	0.139	0.070	0.063	0.025	0.764	0.766	0.738	0.132	0.073		
Mod Thornd		0.418	0.426	0.422	0.421	0.422	0.423	0.920	0.913	0.891	0.423	0.425	0.573	
Rel Mod Thornd		0.408	0.415	0.412	0.379	0.384	0.379	0.990	0.986	0.957	0.414	0.384	0.765	0.931

Notes. Item = item-level personal standard error of measurement (PSEM); Subtest = Subtest-level PSEM; IRT = item response theory SEM; Rel = Jarjoura's SEM; Thornd = Thorndike's SEM; Mod Thornd = modified Thorndike's within-person SEM; Rel Mod Thornd = relative modified Thorndike's within-person SEM.

[a] n = 12,242

Figure Captions

Figure 1. Scatterplots of item-level PSEMs (top three panels) and subtest-level PSEMs (bottom three panels), by IRT model. Open circles show the item-level PSEM at each true score level; open squares show the IRT SEM at each true score level.

Figure 2. Average item-level SEM at each total-correct score by IRT model and type of estimate. Each plot compares four item-level estimates: ITEM = proposed item-level PSEM, IRT = IRT Conditional SEM, LORD = Lord's (1955) SEM, and REL. SEM = Jarjoura's SEM.

Figure 3. Average subtest-level SEM at each total-correct score by IRT model and type of estimate. Each plot compares the four subtest-level estimates. The IRT SEM is plotted for reference. SUB = proposed subtest-level PSEM, THORND = original Thorndike (1951) SEM, MOD.THORND = Thorndike SEM modified to allow subtests of different length, and REL. MOD.THORND = the Thorndike SEM modified to allow subtests of different length and different difficulties.

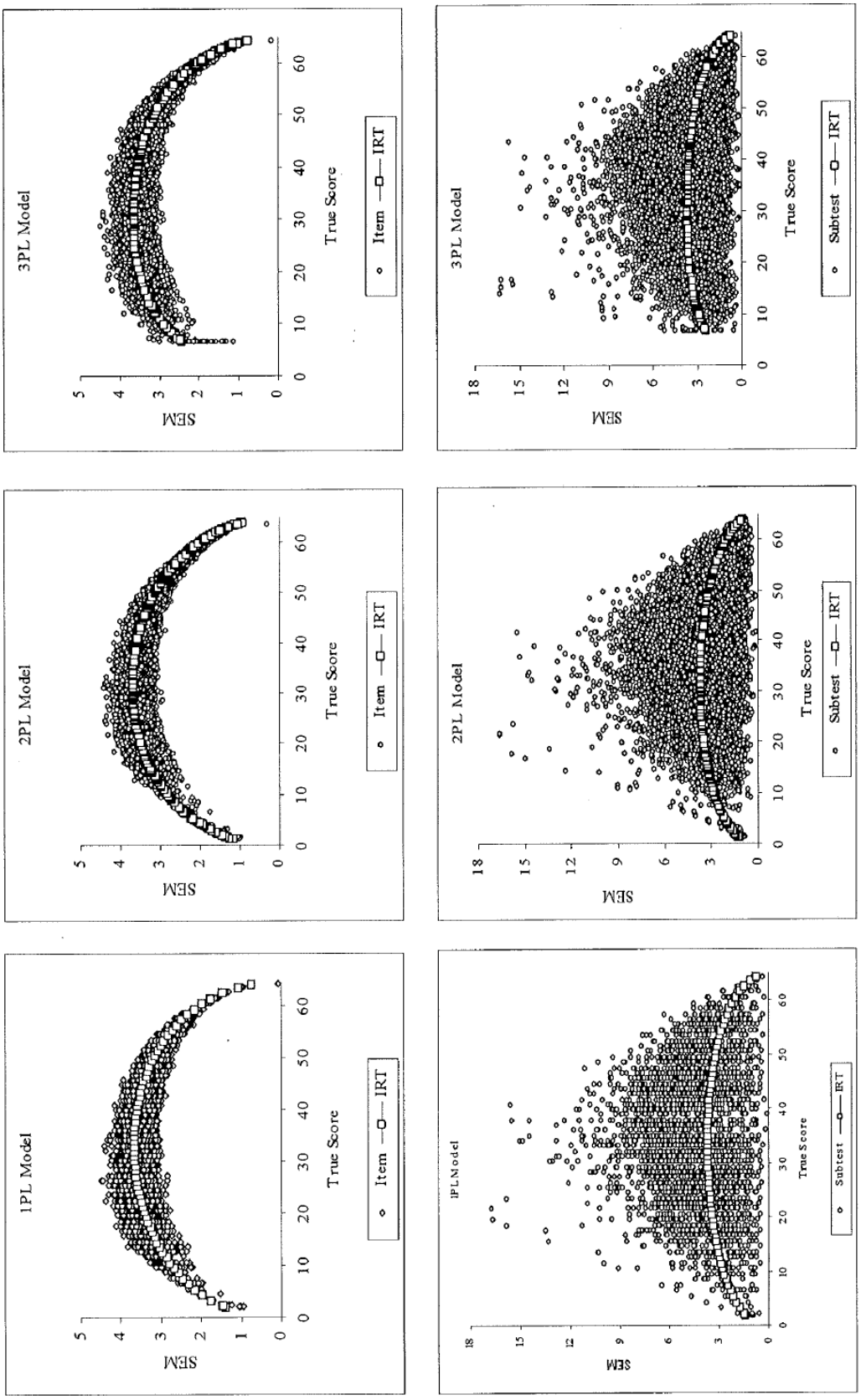
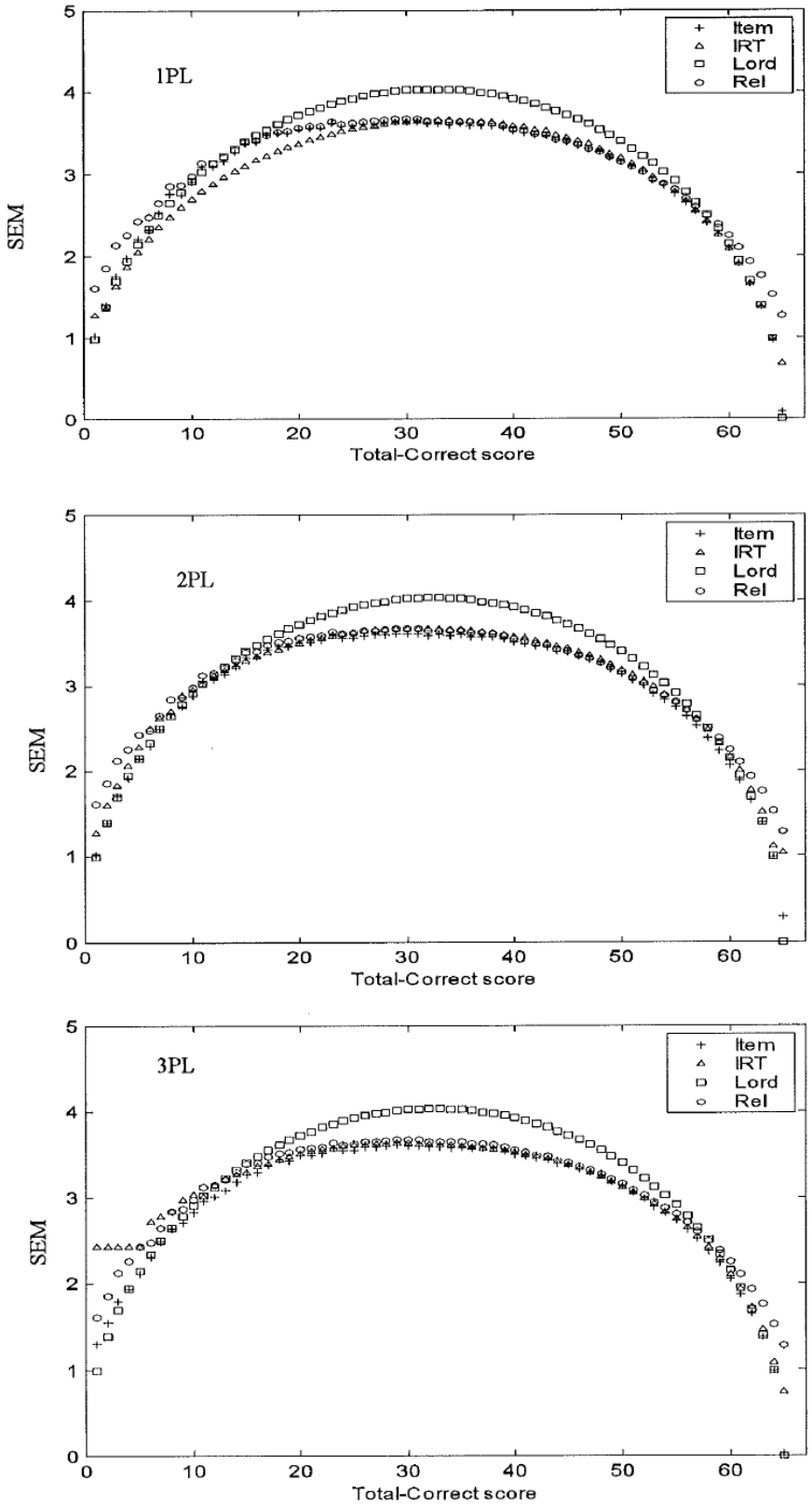


Figure 1. Scatterplots of item-level PSEM, subtest-level PSEM, and IRT SEM at each True Score for the 1PL, 2PL, and 3PL models.

Note. The scatterplot of the IRT SEM in the top and bottom figures are identical for each model.

Figure 2. Average item-level SEM at each total-correct score by IRT model.



Subtest

Figure 3. Average ~~item~~-level SEM at each total-correct score by IRT model.