Using Generalizability Theory to Address Reliability Issues for PARCC Assessments: A White Paper

Robert L. Brennan

Center for Advanced Studies in Measurement and Assessment (CASMA)
University of Iowa

September 9, 2011

Contents

The purpose of this paper is to provide a framework for considering reliability issues for the PARCC assessments. The phrase "reliability issues" covers a number of different statistics that might be estimated, not just reliability coefficients. Indeed, in the author's opinion, reliability coefficients are much less important than other statistics such as standard errors of measurement. The approach taken here uses generalizability (G) theory, particularly multivariate G theory.

G theory offers an extensive conceptual framework and a powerful set of statistical procedures for addressing numerous measurement issues. To an extent, the theory can be viewed as an extension of classical test theory (see, for example, Feldt & Brennan, 1989, and Haertel, 2006) through an application of certain analysis of variance (ANOVA) procedures to measurement issues. Classical theory postulates that an observed score can be decomposed into a "true" score, T, and a single undifferentiated random error term, E. As such, any single application of the classical test theory model cannot clearly differentiate among multiple sources of error. G theory liberalizes classical theory by employing ANOVA methods that allow an investigator to disentangle the multiple sources of error that contribute to the undifferentiated E in classical theory.

In univariate G theory, there is a single universe score (conceptually similar to true score in classical theory). Multivariate G theory, goes a big step further. It considers multiple universe scores and associated composites. As such, multivariate G theory permits designing analyses that closely mirror the structure of complex assessments.

The defining treatment of G theory is a monograph by Cronbach, Gleser, Nanda, and Rajaratnam (1972) entitled *The Dependability of Behavioral Measurements*. Brennan (2001b) provides the most extensive current treatment of G theory. Multivariate G theory was introduced by Cronbach et al. (1972); indeed, they considered multivariate G theory to be the most novel feature of the theory. Brennan (2001b) extended their treatment of multivariate G theory.

This paper is restricted to the two summative components of the PARCC assessments: the machine-scorable (MS) component and the performance-based (PB) component. At the time this white paper was written, the purpose and structure of the two formative components were not sufficiently well defined to address reliability issues in an informed and helpful manner.

1 Illustrative Scenario

To illustrate how multivariate G theory might be used with the PARCC assessments to quantify error variances and measures of precision (e.g., reliability-like coefficients) for the summative components we consider a simplified scenario. This scenario should not be viewed as indicative of what PARCC assessments will look like or should look like. Still, this scenario has features that reflect basic decisions that PARCC has made (e.g., two component assessments that contribute to a summative score) and that illustrate the flexibility of multivariate G theory for addressing issues traditionally associated with the word

"reliability."

Suppose that the MS component of the math assessment consists of items (i) from two different categories m_1 and m_2. It is assumed here that each of these categories will be included in each form of the math MS component, which means that the categories are *fixed* in the terminology of both statistics and G theory. By contrast, the items are assumed to be random in the sense that they differ for each form.[1] Obviously, the assumption of only two categories of items is unrealistic, but the extension to a larger set of categories is straightforward.

Suppose that the PB component of the math assessment consists of three "stations" denoted b_1, b_2, and b_3, where each station involves a different set of math topics or skills. It is assumed here that each of these sets of topics or skills will be included in each form of the math PB component, which means that the stations are *fixed*. By contrast, the actual tasks performed by students at each station are random in the sense that different forms of the PB component involve different tasks. Since the tasks evoke student-produced responses, scores can be obtained only by evaluating these responses. We will assume here that the scoring is performed by human raters. Further, we will assume that each rater evaluates responses to tasks at all stations. (Other possibilities concerning raters are considered briefly later.) This scenario for the PB component well may be oversimplified relative to the assessment design that will be used;[2] nonetheless, this scenario has many features that serve to illustrate the application of multivariate G theory to performance assessments.

The next four subsections consider using multivariate G theory to:

1. conceptualize and analyze the MS component;
2. conceptualize and analyze the PB component;
3. combine results of the MS and PB components; and
4. obtain conditional standard errors of measurement.

The first subsection goes into more detail than the others. To keep the discussion to a manageable length, little consideration is given to estimation issues, which are discussed thoroughly by Brennan (2001b). Also, unless noted otherwise, all discussion is in terms of the mean score metric. So, for example, if all items for the MS component were dichotomously scored, then examinee mean scores would range from 0 to 1, not 0 to the total number of items.[3]

[1] Random sampling assumptions are seldom met in their strictest sense. For purposes here, that is relatively unimportant, and the reader can think of random as "not fixed." What matters here is the distinction between fixed and not fixed.

[2] The PB hypothetical scenario considered here has similarities with some assessments used in medical testing that involve simulated patients.

[3] Converting to the total score metric is not very difficult, but it often causes confusion for those not intimately familiar with G theory.

1.1 MS Component

The illustrative example of the MS component has been studied extensively. It is often called the "table of specifications" model; the basic results are provided by Brennan (2001b, pp. 268–273).

1.1.1 Universe of Admissible Observations

In any multivariate G theory analysis, the building blocks are variance and covariance components for a universe of admissible observations (UAO). For the MS example, the UAO consists of a universe of items for each of the two categories. Strictly speaking, it is assumed that each universe of items is infinite, although it is sufficient for practical purposes that we can conceive of a "large" number of items in each universe. There is also a population of examinees or persons (p) that respond to the items.[4]

The variance and covariance components that characterize the UAO for the MS example are denoted by the elements of the three matrices on the left side of Table 1 ($\mathbf{\Sigma}_p$, $\mathbf{\Sigma}_i$, and $\mathbf{\Sigma}_{pi}$). The σ^2 entries are variance components, and the σ entries are covariance components. These variance and covariance components are interpretable as the results that would be obtained if a type of multivariate analysis of variance were performed on data for the *entire* UAO and population. Estimating the entries in these matrices requires sampling from the UAO and population. The design for doing so is called a G study design. Importantly, whether we are talking about parameters or estimates, the variance and covariance components are for single items, *not* mean scores over some number of items. This "table of specifications" design is often denoted $p^\bullet \times i^\circ$, where:

- $p^\bullet$ signifies that persons are linked across categories—i.e., each person is assessed in all categories; and
- i° signifies that items are independent across categories—i.e., each item appears in one and only one category.

1.1.2 Universe of Generalization and D Study Design

As noted above, the elements of $\mathbf{\Sigma}_p$, $\mathbf{\Sigma}_i$, and $\mathbf{\Sigma}_{pi}$ in Table 1 are for single items. For mean scores over n'_{m_1} and n'_{m_2} items for m_1 and m_2, respectively, $\mathbf{\Sigma}_p$ is unchanged but

$$\mathbf{\Sigma}_I = \begin{bmatrix} \sigma^2_{m_1}(I) & \\ & \sigma^2_{m_2}(I) \end{bmatrix} = \begin{bmatrix} \sigma^2_{m_1}(i)/n'_{m_1} & \\ & \sigma^2_{m_2}(i)/n'_{m_2} \end{bmatrix} \quad (1)$$

and

$$\mathbf{\Sigma}_{pI} = \begin{bmatrix} \sigma^2_{m_1}(pI) & \\ & \sigma^2_{m_2}(pI) \end{bmatrix} = \begin{bmatrix} \sigma^2_{m_1}(pi)/n'_{m_1} & \\ & \sigma^2_{m_2}(pi)/n'_{m_2} \end{bmatrix}, \quad (2)$$

[4]Strictly speaking, it is assumed that the population is infinite, although often the size of the population makes relatively little difference in G theory.

Table 1: Variance-Covariance Matrices for Universes of Admissible Observations

Machine-Scorable		Performance-Based		
m_1	m_2	b_1	b_2	b_3

$$\boldsymbol{\Sigma}_p = \begin{bmatrix} \sigma^2_{m_1}(p) & \sigma_{m_1 m_2}(p) \\ \sigma_{m_2 m_1}(p) & \sigma^2_{m_2}(p) \end{bmatrix} \qquad \boldsymbol{\Sigma}_p = \begin{bmatrix} \sigma^2_{b_1}(p) & & \text{sym} \\ \sigma_{b_2 b_1}(p) & \sigma^2_{b_2}(p) & \\ \sigma_{b_3 b_1}(p) & \sigma_{b_3 b_2}(p) & \sigma^2_{b_3}(p) \end{bmatrix}$$

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} \sigma^2_{b_1}(t) & & \\ & \sigma^2_{b_2}(t) & \\ & & \sigma^2_{b_3}(t) \end{bmatrix}$$

$$\boldsymbol{\Sigma}_r = \begin{bmatrix} \sigma^2_{b_1}(r) & & \text{sym} \\ \sigma_{b_2 b_1}(r) & \sigma^2_{b_2}(r) & \\ \sigma_{b_3 b_1}(r) & \sigma_{b_3 b_2}(r) & \sigma^2_{b_3}(r) \end{bmatrix}$$

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma^2_{m_1}(i) & \\ & \sigma^2_{m_2}(i) \end{bmatrix} \qquad \boldsymbol{\Sigma}_{pt} = \begin{bmatrix} \sigma^2_{b_1}(pt) & & \\ & \sigma^2_{b_2}(pt) & \\ & & \sigma^2_{b_3}(pt) \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{pi} = \begin{bmatrix} \sigma^2_{m_1}(pi) & \\ & \sigma^2_{m_2}(pi) \end{bmatrix} \qquad \boldsymbol{\Sigma}_{pr} = \begin{bmatrix} \sigma^2_{b_1}(pr) & & \text{sym} \\ \sigma_{b_2 b_1}(pr) & \sigma^2_{b_2}(pr) & \\ \sigma_{b_3 b_1}(pr) & \sigma_{b_3 b_2}(pr) & \sigma^2_{b_3}(pr) \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{tr} = \begin{bmatrix} \sigma^2_{b_1}(tr) & & \\ & \sigma^2_{b_2}(tr) & \\ & & \sigma^2_{b_3}(tr) \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{ptr} = \begin{bmatrix} \sigma^2_{b_1}(ptr) & & \\ & \sigma^2_{b_2}(ptr) & \\ & & \sigma^2_{b_3}(ptr) \end{bmatrix}$$

Note. The notation 'sym' in an upper off-diagonal indicates a symmetric matrix.

where I denotes mean scores in the universe of generalization (UG). The associated design is called a D (Decision) study design, which is denoted $p^\bullet \times I^\circ$.

The UG can be conceptualized as an infinite set of randomly parallel forms where all forms have n'_{m_1} and n'_{m_2} items. For any given examinee, τ_1 and τ_2 are the expected values of that examinee's observed scores over forms, where τ_1 and τ_2 are called universe scores. The variance (over persons) of the universe scores are the diagonal elements in $\mathbf{\Sigma}_p$, namely $\sigma^2_{m_1}(p)$ and $\sigma^2_{m_1}(p)$, and the covariance is the off-diagonal element $\sigma_{m_1 m_2}(p)$.

The matrix of universe score variance and covariances components is:

$$\mathbf{\Sigma}_\tau = \mathbf{\Sigma}_p = \begin{bmatrix} \sigma^2_{m_1}(p) & \sigma_{m_1 m_2}(p) \\ \sigma_{m_2 m_1}(p) & \sigma^2_{m_2}(p) \end{bmatrix}. \tag{3}$$

The matrix of relative-error (δ) variance components is:

$$\mathbf{\Sigma}_\delta = \mathbf{\Sigma}_{pI} = \begin{bmatrix} \sigma^2_{m_1}(\delta) & \\ & \sigma^2_{m_2}(\delta) \end{bmatrix} = \begin{bmatrix} \sigma^2_{m_1}(pI) & \\ & \sigma^2_{m_2}(pI) \end{bmatrix}. \tag{4}$$

In this very simple "table of specifications" model, the square roots of the diagonal elements in $\mathbf{\Sigma}_\delta$ are the same as the traditional standard errors of measurement based on coefficient α reliabilities for each of the two categories. The matrix of absolute-error (Δ) variance components is:

$$\begin{aligned} \mathbf{\Sigma}_\Delta &= \mathbf{\Sigma}_I + \mathbf{\Sigma}_{pI} \\ &= \begin{bmatrix} \sigma^2_{m_1}(\Delta) & \\ & \sigma^2_{m_2}(\Delta) \end{bmatrix} = \begin{bmatrix} \sigma^2_{m_1}(I) + \sigma^2_{m_1}(pI) & \\ & \sigma^2_{m_2}(I) + \sigma^2_{m_2}(pI) \end{bmatrix} \end{aligned} \tag{5}$$

1.1.3 Composite for the MS Component

A composite observed mean score for an examinee for the MS assessment can be defined as a weighted mean of the examinee's category mean scores:

$$\overline{X}_{mC} = w_{m_1}\overline{X}_{m_1} + w_{m_2}\overline{X}_{m_2}, \tag{6}$$

where $w_{m_1} + w_{m_2} = 1$, and the person subscript p has been dropped to simplify notation. The w weights are specified a priori by an investigator; they are not the result of some statistical manipulation of the data. Frequently, the w weights are specified in terms of the proportion of items or score points associated with the fixed categories. Equations similar to Equation 6 apply to τ, δ, and Δ.[5] The variances of these composite scores are:

$$\sigma^2_{mC}(\overline{X}) = w^2_{m_1}\,\sigma^2_{m_1}(\overline{X}) + w^2_{m_2}\,\sigma^2_{m_2}(\overline{X}) + 2\,w_{m_1}\,w_{m_2}\,\sigma_{m_1 m_2}(\overline{X}), \tag{7}$$

$$\sigma^2_{mC}(\tau) = w^2_{m_1}\,\sigma^2_{m_1}(\tau) + w^2_{m_2}\,\sigma^2_{m_2}(\tau) + 2\,w_{m_1}\,w_{m_2}\,\sigma_{m_1 m_2}(\tau), \tag{8}$$

$$\sigma^2_{mC}(\delta) = w^2_{m_1}\,\sigma^2_{m_1}(\delta) + w^2_{m_2}\,\sigma^2_{m_2}(\delta), \quad \text{and} \tag{9}$$

$$\sigma^2_{mC}(\Delta) = w^2_{m_1}\,\sigma^2_{m_1}(\Delta) + w^2_{m_2}\,\sigma^2_{m_2}(\Delta). \tag{10}$$

[5] Strictly speaking, the weights used for $\overline{X}$ and τ need not be the same. This issue and its consequences are discussed by Brennan (2001b, p. 307ff).

Given the composite universe score variance and error variances in Equations 8–10, the generalizability coefficient for relative decisions for the MS composite is

$$\boldsymbol{E}\rho^2_{mC} = \frac{\sigma^2_{mC}(\tau)}{\sigma^2_{mC}(\tau) + \sigma^2_{mC}(\delta)}, \tag{11}$$

and the phi coefficient for absolute decisions for the MS composite is

$$\Phi_{mC} = \frac{\sigma^2_{mC}(\tau)}{\sigma^2_{mC}(\tau) + \sigma^2_{mC}(\Delta)}. \tag{12}$$

The reliability-like coefficients are frequently employed, but there are other overall measures of precision. For example, the relative and absolute signal/noise ratios are, respectively,

$$\sigma^2_{mC}(\tau)/\sigma^2_{mC}(\delta) \qquad \text{and} \qquad \sigma^2_{mC}(\tau)/\sigma^2_{mC}(\Delta). \tag{13}$$

Also, assuming tolerance is defined as $\sigma_{mC}(\tau)$, Kane's (1996) error-tolerance ratios for relative and absolute decisions are, respectively,

$$\sigma_{mC}(\delta)/\sigma_{mC}(\tau) \qquad \text{and} \qquad \sigma_{mC}(\Delta)/\sigma_{mC}(\tau). \tag{14}$$

1.2 PB Component

Recall the discussion of the PB component of the math assessment on page 2. The UAO and population are denoted $p^\bullet \times t^\circ \times r^\bullet$, where $\bullet$ and $\circ$ have the interpretations discussed previously, t stands for tasks, and r stands for raters. The variance and covariance component matrices for the UAO are provided on the right side of Table 1. They are 3×3 matrices because there are three fixed stations denoted b_1, b_2, and b_3 in Table 1.

This $p^\bullet \times t^\circ \times r^\bullet$ design is obviously a more complicated situation than the "table of specifications" design for the MS component largely because of the inclusion of both a task facet and a rater facet. Since each rater rates responses for examinees at all stations, the matrix $\boldsymbol{\Sigma}_r$ is a full symmetric matrix, where the off-diagonal elements are covariance components. (In a situation like this, we often say that the rater facet is linked.) The $\boldsymbol{\Sigma}_{pr}$ matrix is also a full symmetric matrix because each person-rater combination contributes data to all three stations.

The logic and steps outlined in Section 1.1 for the MS component can be extended to the design for the PB component. Perhaps the most important difference between the MS and PB components is that the UG for the MS component contains a single random item facet (I), whereas the UG for the PB component contains both a task (T) facet and a rater (R) facet, both of which are random. All other things being equal, as the number of random facets in the UG increases, error variances increase too.

The D study design for the PB component is $p^\bullet \times T^\circ \times R^\bullet$, and the universe score, relative error, and absolute error matrices are:

$$\boldsymbol{\Sigma}_\tau = \begin{bmatrix} \sigma^2_{b_1}(p) & & \text{sym} \\ \sigma_{b_2 b_1}(p) & \sigma^2_{b_2}(p) & \\ \sigma_{b_3 b_1}(p) & \sigma_{b_3 b_2}(p) & \sigma^2_{b_3}(p) \end{bmatrix} \tag{15}$$

$$\boldsymbol{\Sigma}_\delta = \begin{bmatrix} \sigma^2_{b_1}(\delta) & & \text{sym} \\ \sigma_{b_2 b_1}(\delta) & \sigma^2_{b_2}(\delta) & \\ \sigma_{b_3 b_1}(\delta) & \sigma_{b_3 b_2}(\delta) & \sigma^2_{b_3}(\delta) \end{bmatrix} \tag{16}$$

$$\boldsymbol{\Sigma}_\Delta = \begin{bmatrix} \sigma^2_{b_1}(\Delta) & & \text{sym} \\ \sigma_{b_2 b_1}(\Delta) & \sigma^2_{b_2}(\Delta) & \\ \sigma_{b_3 b_1}(\Delta) & \sigma_{b_3 b_2}(\Delta) & \sigma^2_{b_3}(\Delta) \end{bmatrix}. \tag{17}$$

Note that the off-diagonal elements (covariance components) of $\boldsymbol{\Sigma}_\delta$ and $\boldsymbol{\Sigma}_\Delta$ are *not* 0, as they are for the MS component (see Equations 4 and 5). This means that both $\boldsymbol{\Sigma}_\delta$ and $\boldsymbol{\Sigma}_\Delta$ may involve correlated error. In this hypothetical scenario, correlated error might arise, for example, as a result of rater halo effects across stations.

The incorporation of correlated error in multivariate G theory is a distinguishing and important feature of the theory. Indeed, correlated error arises naturally in multivariate G theory, rather than being an abstract "add on" that involves additional assumptions. It is particularly important to note that since covariance components can be either positive or negative, correlated error either increase or decrease the composite error variance, which is briefly discussed next.

A composite observed mean score for an examinee for the PB assessment can be defined as a weighted mean of the examinee's category mean scores:

$$\overline{X}_{bC} = w_{b_1}\overline{X}_{b_1} + w_{b_2}\overline{X}_{b_2} + w_{b_3}\overline{X}_{b_3}, \tag{18}$$

where the sum of the w weights is set to 1 to give the composite a mean-score metric interpretation. Composites for τ_{bC}, δ_{bC}, and Δ_{bC} can be defined similarly. Their variances can be obtained in the manner discussed in Section 1.1.3. Reliability-like coefficients and other measures of precision can be computed using Equations 11–14 (with obvious changes in subscripts).

1.3 MS and PB Composite

Equations 6 and 18 for the MS and PB observed-score composite are repeated below:

$$\overline{X}_{mC} = w_{m_1}\overline{X}_{m_1} + w_{m_2}\overline{X}_{m_2},$$

and

$$\overline{X}_{bC} = w_{b_1}\overline{X}_{b_1} + w_{b_2}\overline{X}_{b_2} + w_{b_3}\overline{X}_{b_3}.$$

Recall that the w weights are specified a priori by an investigator; they are not the result of some statistical manipulation of the data.

To obtain an overall observed-score composite for the MS and PB components, let v_m be the a priori weight for $\overline{X}_{mC}$ and let v_b be the a priori weight for $\overline{X}_{bC}$. Then, the overall observed-score composite is:

$$\overline{X}_C \quad = \quad v_m \, \overline{X}_{mC} + v_b \, \overline{X}_{bC}. \tag{19}$$

$\sigma^2_C(\overline{X})$ can be obtained by using Equation 19 to compute the overall composite for each examinee and then obtaining the variance.

Obtaining the overall composite error variances is straightforward. Since the composite error variances for MS and PB are uncorrelated,

$$\sigma^2_C(\delta) = v^2_m \, \sigma^2_{mC}(\delta) + v^2_b \, \sigma^2_{bC}(\delta) \tag{20}$$

and

$$\sigma^2_C(\Delta) = v^2_m \, \sigma^2_{mC}(\Delta) + v^2_b \, \sigma^2_{bC}(\Delta). \tag{21}$$

Finally, since $\sigma^2_C(\overline{X}) = \sigma^2_C(\tau) + \sigma^2_C(\delta)$, the overall composite universe score variance is

$$\sigma^2_C(\tau) = \sigma^2_C(\overline{X}) - \sigma^2_C(\delta). \tag{22}$$

Given these results, reliability-like coefficients and other measures of precision can be computed using Equations 11–14 (with obvious changes in subscripts).

1.4 Conditional Standard Errors of Measurement

Conditional standard errors of measurement (CSEMs) are standard errors of measurement for individual persons. The basic idea is to obtain the variance of the mean (or weighted mean) based on the data for person p only. Multivariate G theory provides a rather natural way to do so. Brennan (1998) discusses this in detail; Brennan (2001b, pp. 314–317) provides a briefer treatment.

Consider the MS component. For any given person, the data consist solely of item scores for the two categories. This rather trivial design can be denoted I°. Let X_{pi1} be item scores for person p for m_1, and let X_{pi2} be item scores for person p for m_2. Then the absolute CSEM for person p is

$$\sigma_{mC}(\Delta_p) = \sqrt{w^2_{m_1} \frac{\sigma^2(X_{pi1})}{n'_{m_1}} + w^2_{m_2} \frac{\sigma^2(X_{pi2})}{n'_{m_2}}},$$

where the variances are taken over item scores within categories.[6]

For the PB component, the data for a particular person can be denoted $T^\circ \times R^\bullet$. Computing $\sigma_{bC}(\Delta_p)$ involves several steps that are very similar to the example in Brennan (2001b, pp. 316–317).

Since errors are uncorrelated across MS and PB, the overall CSEM for person p is

$$\sigma_C(\Delta_p) = \sqrt{v^2_m \, \sigma^2_{mC}(\Delta_p) + v^2_b \, \sigma^2_{bC}(\Delta_p)}.$$

[6]When all items are dichotomously-scored, this CSEM is the square root of Equation 49 in Feldt and Brennnan (1989, p. 124.)

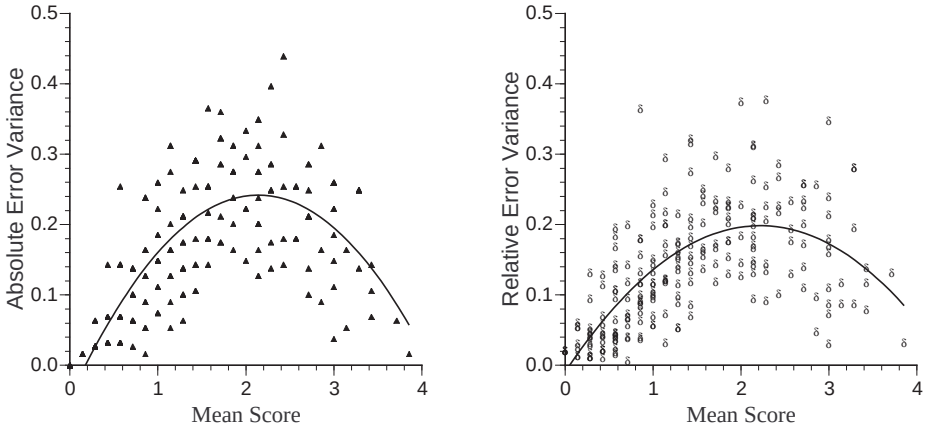

Figure 1: An example of conditional error variances.

The above brief discussion of CSEMs does not explicitly consider the fact that we usually want CSEMs for one or more reported scale scores. That matter is briefly considered in Section 2.1.

Although the CSEMs discussed above are for individual examinees, strictly speaking CSEMs are needed only for scale scores that might be achieved. For most scales used for reporting, the number of scale score points, say k, is much smaller than the number of examinees tested. This might appear to suggest that we need to compute CSEMs for only k examinees. This line of reasoning, however, is flawed in that it is blind to the fact that almost always CSEMs for examinees with the same scale score tend to be quite variable.[7] This is illustrated by Figure 1 which is taken from Brennan (2001b, p. 163).

The data that generated the results in Figure 1 are for a design that is much simpler, and an assessment that is likely much shorter, than the PB component design ($p^\bullet \times T^\circ \times R^\bullet$). Still the figure illustrates the variability of estimated CSEMs for different persons with the same mean score, although the variability may be exaggerated relative to what will be observed with the PB component. For a given mean score, the variability in CSEMs is partly attributable to noise (i.e., random error) and partly attributable to the fact that the same mean score may arise from different patterns of item, task, and/or rater scores. In many practical contexts, it is judged unacceptable to provide different estimated CSEMs to examinees with the same reported score. Accordingly, it usually makes sense to smooth the results in some manner. A quadratic fit such as that in Figure 1 is often adequate for practical use (see Brennan, 2001b, pp. 162–164).

The left sub-figure in Figure 1 is for absolute error CSEMs; the right sub-figure is for relative error CSEMs. Formulas for relative error CSEMs are discussed by Brennan (2001b, pp. 161–162), although they are not nearly as fre-

[7]See the subsequent discussion of standard errors in Section 2.4.

quently used as are the absolute error CSEMs.

It is noteworthy that the two sub-figures have a concave down quadratic fit. By contrast, IRT CSEMs almost always display a concave up pattern. As discussed by Brennan (1998), this is attributable to the fact that the θ scale involves a severe stretching of the raw-score scale, which is the scale used in Figure 1. Obviously, the scale that should be used is the one employed for reporting assessment results, which is seldom either the raw-score scale or the theta scale. In short, the choice of scale matters not only for reporting purposes but also for quantifying CSEMs.

The most common practice is to first pick a score scale and then accept whatever the CSEMs happen to be. An alternative procedure is to first specify one or more desired characteristics of the CSEMs, and then accept whatever this implies for the score scale. The latter approach is basically what was done in the late 1980s when the ACT Assessment was last rescaled. (See Brennan, 1989, for a full discussion of the methodology and results.) In that context, it was decided to make the CSEMs as constant as possible throughout the full scale-score range. This was done using an arcsine transformation of raw scores (see Kolen, 1988, Kolen & Hanson, 1989, and Kolen & Brennan, 2004, pp. 348ff.)

2 Brief Consideration of Other Issues

This section provides brief treatments of a number of other issues that undoubtedly will arise in considering error variances and measurement precision issues for PARCC. The discussions here are far from complete. Although, it is known how to deal with many of these issues, in general, it is almost certain that considerable work will need to be done (i.e., a program of research) to deal with these issues in the PARCC context, once program design decisions are made. In this section, the personalized "I" is sometimes used rather than "we" when I want to emphasize that some statement is my opinion, as opposed to a collective judgement.

2.1 Converting Results to Reported Score Scales

Multivariate G theory analyses are performed with observed scores (usually in the mean-score metric, as discussed previously). In many situations, however, final results (e.g., error variances) need to be converted to one or more reported score scales (RSSs). For a particular RSS, let S(X) designate the transformation of observed scores to the RSS. If S(X) is a linear transformation of raw scores, obtaining quantities such as transformed error variances is straightforward. In most cases, however, S(X) a non-linear transformation of raw scores.

2.1.1 Non-Linear Transformations

For a non-linear transformation, it is easy to get the variance of the transformed observed scores—simply use S(X) directly to get each examinee's scale score,

and then take the variance. Obtaining the error variance for non-linearly transformed raw scores is more complicated, but this matter has been the subject of considerable research in the past 20 years (see, for example, Kolen, Hanson, & Brennan, 1992, Brennan & Lee (1999) and Lee, Brennan, & Kolen, 2000). I think it is safe to say that the field now has sufficiently good procedures to deal with non-linear transformations of raw scores.

2.1.2 Equating

In testing programs with multiple forms (which is surely the case for PARCC), raw scores for all forms need to be put on a common scale. Under traditional equating procedures, scale scores are defined for the initial form. Then, raw scores for subsequent forms are equated to raw scores for the initial form, and finally these equated raw scores are converted to scale scores. Under IRT, there are two types of equating procedures: IRT true score equating, and IRT observed score equating. The latter is reasonably similar to traditional equating procedures, the former is conceptually and analytically rather different. In all cases, however, equating is properly viewed as a process to maintain the score scale over multiple forms.

For PARCC, however, equating may be quite challenging in that equating tends to become suspect as the populations who take different forms become more dissimilar. It may be safe to assume that the populations of examinees who choose to test within a window are not too disparate; the same assumption, however, may be quite untenable across windows. Dissimilarity in populations is often dramatically true for the first few years of a new testing program. This is one of several reasons why it would be prudent to plan for a rescaling of PARCC assessments after, say, two or three years.

2.2 Analyses by Window and Across Windows

Currently, I do not know for certain whether or not there will be multiple forms of the summative assessments in a given window. If there are multiple forms, the best way to deal with them within a window is to obtain averages over forms of the estimated variance and covariance components (e.g., those entries in the matrices in Table 1). Then, the types of procedures discussed in Section 1.1 can be used to obtain results over forms. In doing so, this means that there are *not* form-specific estimates of quantities such as error variances. This is not a limitation; rather, it is entirely consistent with the assumptions of G theory. Each form is viewed as one of the forms in the UG, with the parameters being expected values *over* forms. The best estimates of these parameters involve estimates over all available forms in the window. Furthermore, the variabilities of such estimates are empirically-based estimates of standard errors, which can help considerably in determining how much confidence to have in the results.

How to deal with different windows is an open question until we know more about them. For example, if the time-frames for the windows are not too different, then the averaging procedure discussed above is probably sensible. More

likely, however, the windows will be different enough in time and/or the types of examinees involved that averaging may not be sensible, and separate results for each window (for both the MS and PB components) will be required. One crucial issue (and perhaps the most crucial issue) will be whether or not it is reasonable to expect that substantial learning likely could occur during the time span of the various windows.

2.3 Disaggregation

It is likely that there will be a need to obtain error variances and measures of precision for different subgroups of the total group. The obvious way to accommodate this need is to conduct analyses for each subgroup separately. Doing so, will generate substantial computer output, but the process is very straightforward. Sometimes, however, results are misinterpreted.

For example, at the risk of oversimplification, error variances for examinees in various subgroups often are not much different. In fact, a CSEM is completely blind to any characteristics of the individual person, except his or her scores. What usually differs (and sometimes substantially) is universe score variance, which changes reliability-like coefficients, even if error variances are largely unchanged. When subgroup "reliability" results are reported publicly, often they are misleadingly interpreted as meaning that reliability for one subgroup is a lot different from another subgroup. Such statements may be factually correct in a narrow sense but, in my opinion, the most important results are captured largely by the CSEMs, not reliability-like coefficients.

2.4 Sample Sizes and Standard Errors

Traditionally, in the behavioral and social sciences, sample size issues are discussed in the generic context of standard errors where, almost always, the standard errors are for statistics based on sampling persons. In G theory, sampling of persons is relevant, but persons are certainly not the sole "facet" of interest, and often they are not even the most important facet. Consider, for example, the PB component, as discussed in Section 1.2. Letting persons be called a facet, there are three facets: persons, tasks, and raters. We do not include type of task as a facet, because it is fixed (i.e, the same three types of tasks reoccur for every replication).

As far as G theory is concerned, the crucial questions that relate to sample sizes are:

1. what are the estimated standard errors of the estimated variance components (and covariance components, if any) and D study statistics such as coefficients and error variances; and

2. what sample sizes are needed to make these estimates (especially the D study estimates) tolerably small?

Answering the first question is complicated, involving numerous statistical issues and many possible procedures (see, for example, Brennan, 2001b, chap. 6, and pp. 249–251, 328–330).

A thorough answer to the second question is also complicated, but in practice, it is not too difficult to provide reasonable guidance. First, for any subgroup (including the total group) experience suggests that about 1000 persons is usually adequate, and only a few hundred may be sufficient. Often, the issue that is more crucial than sample size is the representativeness of the sample relative to the population of interest.

Second, for facets other than persons, practical constraints such as cost and testing time often put a very strict upper limit on sample sizes. Still, larger sample sizes are obviously better than smaller sample sizes. Beyond that, the following guidance based my experience may be helpful:

- as discussed more fully in the next subsection, at least two levels of each facet are necessary to avoid ambiguities caused by confounding;
- almost always it is best to have as many tasks as possible, consistent with time constraints; and
- two raters are often sufficient if they are well trained and rubrics are carefully constructed and conscientiously followed.

2.5 Other Facets and Complexities

The illustrative scenarios discussed in Section 1 are reasonably realistic but they are still probably over-simplified relative to the ultimate design of the PARCC assessments. For example, for purposes of simplicity, it is assumed in Table 1 that for the PB component each rater rates the responses to all tasks at all stations for all examinees. This is not likely to be the design ultimately chosen by PARCC. It is much more likely that different sets of examinees will be rated by different sets of raters and perhaps any given rater will rate responses for only a subset of the stations. Such designs may be logistically easier to implement, less time-consuming, and/or less costly, but they almost always add complexity to G theory analyses. For this reason, it is important that decisions about the assignment of raters to examinees and tasks be made jointly by those responsible for the ratings and those responsible for psychometric analyses. If this is not done, it can be very difficult (and sometimes impossible) to get acceptably accurate estimates of the contribution of raters to error variance.

Also, facets other than those mentioned might well be involved. For example, in ELA, passages or writing exercises might be involved, and for ELA and/or math raters might be augmented or replaced by an automated scoring engine. These types of facets can introduce considerable complexities. Also, serious consideration of any measurement procedure almost always leads to the conclusion that occasion is an intended facet in the UG, but this facet is often hidden in a D study. These issues are discussed more fully next.

2.5.1 Confounded Effects and the "Problem of One"

Consider an ELA high-stakes testing program that involves reading passages. Almost always for security purposes different passages are used for each form, which means that the passages are random from the perspective of G theory. Furthermore, almost always for the various passages, questions or items are associated with fixed content and/or skill categories in a table of specifications, and there are different numbers of items associated with each cell in the table (i.e., the design is unbalanced).

In addition, it is not uncommon that there is only one passage for each of some small number of types of passages. In a similar vein, if the ELA assessment involves writing exercises, often there is only one exercise for each of several types of writing. This "only one" issue looms large, as discussed below.

Suppose an ELA assessment involves two writing exercises, one narrative and one persuasive. Suppose, as well, that the writing exercises (z) are different over forms, but the types of exercises (t) are the same over forms, narrative and persuasive. This means that:

- z is random, but t is fixed in the universe; and
- z and t are completely confounded because there is only one exercise for each type of exercise.

We denote this confounding as (z, t). In this example, the confounding has two related problematic consequences:

- any G or D study that involves (z, t) cannot distinguish effects attributable to z from effects attributable to t; and
- any G or D study that involves (z, t) must treat (z, t) as either random or fixed, but either choice is "half" wrong, since z is random and t is fixed in the universe.

The net consequence is that, unless something is done to resolve the confounding, estimates of error variance and measures of precision will be biased in all but trivial situations; and the best that can be done is to provide some range (often unacceptably large) of possible values for the statistics of interest.

Resolving such confounding is possible, but it requires a side study that might be performed in a pilot or field trial. The key issue in conducting such a study is that there must be at least two writing exercises for each type of writing. Only then can writing effects be distinguished from type-of-writing effects. Designing such a study can be challenging, but it is indispensable to resolving confounding.

It is possible, of course, that such a side study could reveal that the type of writing makes no difference with respect to student performance. That is, an analysis could reveal that results are invariant with respect to type of writing. If so, the type-of-writing facet could be dropped from the universe, which resolves the psychometric problem, provided doing so is acceptable to those responsible for assessment design.

An important "take-home" message from the above discussion is that any psychometric analysis with extant data (not just G theory) is likely to get into trouble if there is only one of something, because a single condition of a facet cannot provide any evidence of variance over conditions of the facet.

2.5.2 Automated Scoring Engines and the "Problem of One"

It is unclear to me when/if an automated scoring engine(ASE) will be used with the summative PARCC assessments. However, let's suppose that an ASE is used eventually. Broadly speaking, there are two possible types of use: (a) an ASE replaces all raters; and (b) an ASE replaces some raters. The latter possibility raises all sorts of complicated design issues, as well as reliability issues—so much so that I am not going to attempt to address (b) here.

If an ASE replaces all raters we again have the "problem of one"—i.e., any particular ASE is essentially a single computer algorithm. There is no reason to believe that one such algorithm (i.e., a particular ASE) would necessarily give the same rating results as another algorithm (i.e., another ASE). Indeed, each company that has an ASE argues vehemently that their ASE is better in some sense than some other company's ASE. Data to support such arguments are often scant or non-existent, but the arguments persist. Assuming PARCC picks one ASE, then the ASE facet (which essentially plays the same role as the rater facet) will be confounded with at least the b fixed facet and perhaps both b and the random t facet. Under such circumstances, with operational data, only, it will be impossible to distinguish between effects attributable to the particular ASE chosen by PARCC and at least some other effects.

This conundrum could be addressed with a side study using at least two ASEs. This would be the best approach, I think, but probably it would be very difficult to implement. In the absence of such a study, there will need to be other side studies that address at least two issues:

- the variability of ASE ratings using different sets of training papers, and
- the variability of ratings using the particular ASE vis-a-vis using human raters.

These studies will need to be carefully designed and thoughtfully analyzed so that the variability of ratings is incorporated into error variance for reported examinee scores.

Frequently some evidence is provided that an ASE gives scores that are comparable to those obtained with human raters. Providing such evidence is reasonable, but it implies that scores obtained using the ASE are no more reliable than human raters; i.e., the ASE ratings have some degree of unreliability. Unfortunately, too often this unreliability is *not* incorporated into error variances for reported examinee scores, as it should be.

2.5.3 Occasion as a Facet and the "Problem of One"

From the perspective of G theory, various types of reliability coefficients and error variances relate conceptually to consistencies and inconsistencies, respectively, over *replications* of a measurement procedure (see Brennan, 2001a). This forces the investigator to grapple with both the conceptional issue and the operational definition of what constitutes a replication of the measurement procedure. Almost always, serious consideration of this matter leads to the conclusion that replications are intended to involve generalizing over some set of occasions. A common problem, however, is that the only available data are for an operational assessment (i.e., D study) in which each examinee takes the assessment on a *single* occasion, which means that occasion is a hidden, fixed facet.[8]

In such a case, estimates of reliabilities will be inflated, and estimates of error variances will be deflated. This problem can be resolved with a side study in which a representative group of examinees takes two different forms of an assessment at two different points in time.[9] The two occasions need to be chosen carefully so that the investigator has reasonable confidence that examinees' universe scores have not changed much over the two occasions.

The occasion facet can play a frequently overlooked role in PB assessments. For such assessments, it it well known that estimates of the person-task interaction variance component, $\sigma^2(pt)$, tend to be quite large which, in most cases, leads to large error variances. Almost always, however, PB assessments involve a single, fixed occasion facet. Under such circumstances, there is some fairly compelling evidence that the size of the estimate of $\sigma^2(pt)$ is often largely attributable to $\sigma^2(pto)$ (see, for example, Cronbach, Linn, Brennan, & Haertel, 1997), but estimates of $\sigma^2(pt)$ cannot be distinguished from estimates of $\sigma^2(pto)$ without a side study that involves at least two occasions.

The messages here are quite clear. First, failing to incorporate occasion as a facet in an operational assessment, when occasion is an intended facet in the UG, will likely lead to overestimating reliabilities, underestimating error variances, and misunderstanding the relative contribution of facets to various error variances. Second, a program of research that includes side studies with at least two occasions can help considerably with disentangling the confounded effects caused by use of a single occasion in an operational assessment.

2.6 Reliability of Growth Measures

It seems likely that PARCC will report some kind of measure(s) of growth. However, to the best of my knowledge the subject of error variances and measures

[8] In an operational setting, sometimes different groups of examinees are tested using the same assessment on thinly separated occasions. Then separate estimates of reliabilities and error variances are computed for each occasion (i.e., set of examinees), and the results are averaged in some manner. This process does *not* involve generalizing over occasions; rather, this process is merely averaging results for different *fixed* occasions using the same assessment.

[9] If the group is of adequate size, it may be preferable to randomly split it in half, and have the forms administered in a counterbalanced manner (i.e., one subgroup takes the first form followed by the second, and the other subgroup takes the second form followed by the first).

of precision for measures of growth is largely uncharted territory. The topic is closely related to measures of change, or modified measures of change (see Feldt & Brennan, 1989, pp. 118–120, and Haertel, 2006, pp. 79–80), which has been a topic of unending debate for at least 50 years. I am relatively optimistic that once PARCC decides which measure(s) of growth are going to be used, it will be possible to address error variance and measurement precision issues in a reasonable manner. Doing so will require some challenging research, however, which likely will not resolve all issues associated with the long-standing debate about measures of change.

2.7 G theory and IRT

I have spent considerable time during the past decade attempting to integrate G theory and IRT. For a number of reasons, I no longer believe such an integration is possible given the current conceptions and assumptions of the two theories. Two (of many) issues are briefly considered below to illustrate my claim.

- IRT, as the name suggests, focuses on items as a starting point, and it can be a powerful tool for test construction in many contexts. By contrast G theory focuses on *collections* of items and other facets; G theory has nothing to say about individual items, or individual conditions of any other facet. In this sense, IRT can be viewed as a micro theory and G theory can be viewed as a macro theory.
- Perhaps the most central issue in G theory is the identification of measurement facets and deciding which of them are to be viewed as random and which fixed. By contrast, IRT cannot handle multiple random facets simultaneously. There have been minimally successful attempts to do so under restricted circumstances (e.g., Bock, Brennan, & Muraki, 2000), but that is all, thus far.

That does not mean, however, that IRT and G theory cannot complement each other in real-life testing contexts, provided investigators are thoughtfully pragmatic without being irresponsible. I have little doubt, for example, that IRT can be used successfully for developing PARCC assessments, but doing so is going to require the development of novel items and careful consideration of different pools of items that assess different standards or sets of standards from the CCSS. Similarly, I have little doubt that G theory can be used successfully to address issues traditionally associated with the word "reliability," but doing so will require thoughtful decisions about facets and assessment design, as well as non-traditional and likely complex analyses.[10]

2.8 Estimation Issues and Computer Programs

Brennan (2001b) treats estimation issues in G theory in considerable detail. Note, in particular, that Chapters 9 and 10 in Brennan (2001b) consider the

[10]To put it bluntly, coefficient α is not likely ever to be a defensible statistic for characterizing the reliability of scores for a PARCC assessment.

two designs in Table 1.

The CASMA website (www.education.uiowa.edu/centers/casma) has a suite of free computer programs (and associated manuals) for performing generalizability analyses: GENOVA (Crick & Brennan, 1983) for univariate G and D studies with balanced designs; urGENOVA (Brennan, 2001d) for univariate G studies with unbalanced designs; and mGENOVA (Brennan, 2001c) for a selected set of multivariate G and D study designs. These programs are briefly described in Appendices F, G, and H in Brennan (2001b). The two designs in Table 1 can be analyzed using mGENOVA.

Sometimes, other compute packages can be used to estimate some G theory results, but care must be exercised, particularly if one or more facets is/are fixed, and/or designs are unbalanced. Furthermore, all such packages have considerable difficulty with the large data sets that characterize the types of analyses that PARCC is likely to want to conduct. Virtually all general purpose statistical packages for estimating the variance components used in G theory do so using matrix procedures that involve manipulating matrices that have as many (or nearly as many) rows and columns as there are observations in the data set. These matters are discussed rather extensively by Brennan (2001b, see especially pp. 241–247).

Note that none of the best known statistical packages can handle multivariate generalizability analyses (which involve estimated covariance components as well as variance components), except in the sense that packages such as SAS IML or R could be used to program analyses "from scratch," provided large data sets can be accommodated.

2.9 Need for Multivariate G Theory

G theory provides an investigator with a conceptual framework and statistical methodology for disentangling the multiple facets that are confounded in the single error term of classical theory. This permits G theory to answer questions about the contribution of different random facets to error variances and coefficients, which cannot be done with classical theory or IRT. These statements apply to both univariate and multivariate G theory. For PARCC, however, multivariate G theory seems much more relevant.

Consider, again, the PB illustrative scenario in Table 1. The multivariate G theory analysis that gives the PB results in Table 1 is essentially three univariate analyses that give the diagonal elements in the variance-covariance matrices on the right side of Table 1, combined with three analyses that give the covariance components in the off diagonals. It is relatively straightforward to get the variance components and usually not much more difficult to get the covariance components.

In principal, it is nearly always possible to perform a univariate G theory analysis that imperfectly mirrors a multivariate analysis. Basically, the univariate analysis collapses over levels of the fixed facet. In doing so, however, a collapsed univariate analysis

- hides the covariance components, which can be very informative (recall the previous discussion of correlated error);
- hides the similarities, differences, and relative contributions of the various levels of the fixed facet;
- cannot accommodate a priori weights for a composite (except weights that are proportional to sample sizes); and
- is almost always more difficult to understand and perform than a multivariate analysis.

The last comment may seem strange, but it is true in my experience.

Perhaps most importantly, a multivariate analysis of the PB component is almost certainly a better modeling of the assessment than a univariate analysis that involves collapsing over levels of the fixed facet.

2.10 Field Tests vs. Operational Administrations

For the overall PARCC summative assessments, the fundamental statistics that need to be estimated are variance and covariance components like those in Table 1 for the illustrative scenario. Given these statistics, all other results for the overall assessments can be estimated. For CSEMs, separate analyses are required for individual examinees.

Ideally, data should be collected in a field test for at least a couple of forms that have all the features of the operational assessments. Then, analyses of the type described in Section 1 can be performed to decide whether revisions in the assessment design are needed. To the extent possible, however, final estimates of reliability and error variance should rely on operational data from the live windows.

As discussed in Section 2.5, depending on how the operational assessments are conceptualized and designed, it may be necessary to use field tests for side studies that permit estimating certain parameters that cannot be estimated with the operational assessments.

Sometimes, for practical reasons, forms of the full operational assessments cannot be administered in their entirety to a representative sample of examinees in a field test. These situations present formidable challenges because the field test must be designed in such a way that all necessary parameters are estimated, which usually means employing different but randomly equivalent groups of examinees.

3 References

Brennan, R. L. (Ed.). (1989). *Methodology used in scaling the ACT Assessment and PACT+*. Iowa City, IA: American College Testing.

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

Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295–317.

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

Brennan, R. L. (2001c). *mGENOVA (Version 2.1)* [Computer software and manual]. Iowa City, IA: University of Iowa. (Available on http://www.education.uiowa.edu/centers/casma)

Brennan, R. L. (2001d). *urGENOVA (Version 2.1)* [Computer software and manual]. Iowa City, IA: University of Iowa. (Available on http://www.education.uiowa.edu/centers/casma)

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

Crick, J. E., & Brennan, R. L. (1983). *GENOVA: A generalized analysis of variance system* [Computer software and manual]. Iowa City, IA: University of Iowa. (Available on http://www.education.uiowa.edu/centers/casma/)

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373–399.

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 and Macmillan.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (1996). The precision of measurements. *Applied Measurement in Education*, *9*, 355–379.

Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement, 25*, 97–110.

Kolen, M. J. & Hanson, B. A. (1989). Scaling the ACT Assessment. In R. L. Brennan (Ed.), *Methodology used in scaling the ACT Assessment and PACT+* (pp. 35–55). Iowa City, IA: American College Testing.

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.

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37*, 1–20.