Overview of C-SEM Methods
Daniel Biddle, Ph.D.
Biddle Consulting Group

The traditional Standard Error of Measurement (SEM)\(^1\) should no longer be used in practice unless the researcher is confident that measurement error is consistent throughout the score range—a circumstance that is rare indeed\(^2\). Rather, the *conditional* SEM should be used because it accurately reflects the *variation of measurement precision* throughout the range of scores.

The fact that measurement precision (i.e., represented in SEM units) varies throughout the score range was allegedly first discovered by William G. Mollenkopf, a professor at Princeton University, while studying under the prestigious Harold Gulliksen (author of *The Theory of Mental Tests* in 1950)\(^3\). Mollenkopf’s studies resulted in a 1949 publication in Psychometrika titled, “Variation of the standard error of measurement of scores” and was subsequently followed by dozens of articles that explored the same phenomena.

Leading up to the last decade, the most significant milestone for the C-SEM research thread occurred with the publication of the 1999 *Standards for Educational and Psychological Testing*, where—for the first time in measurement history—the *conditional* SEM was codified into APA standards (Standard 2.2 and 2.14). The conditional SEM has been regularly used in the educational measurement community since its inception, however, the concept only began showing up in I-O psychology publications in more recent years (e.g., Bobko & Roth, 2004; Biddle, 2005, 2008, Biddle *et al.*, 2007).

More than a dozen various C-SEM techniques have accumulated since the first mention of the concept over 58 years ago. Some techniques are original; others are “improved” techniques that include minor adjustments to the original author’s work. Some techniques are based on “strong true score models,” which are tied to strong statistical and theoretical (i.e., parametric) assumptions. Binomial error models and IRT models for computing C-SEMs generally fall into this category. Other C-SEM models are categorized as “weak true score models” and encompass the “difference methods” that were used by the original C-SEM authors. Such methods are referred to as the “difference” methods because they compute the C-SEM based on the variance between two split halves of the same test. These methods are referred to as the “weak” methods because they are not fundamentally tied to theoretical or parametric distributional assumptions.

While the techniques for computing C-SEMs are widely varied, they are fortunately rather close in their outputs. For example, Qualls-Payne (1992) showed that the average C- SEM produced by the difference method (across 13 different score levels) produced C-SEM values that were very close to five other methods (including binomial

---

\(^{1}\) \(SEM = \sigma_x (1 - r_{xx})^{1/2}\) where \(\sigma_x\) is the standard deviation and \(r_{xx}\) is the reliability of the test.

\(^{2}\) The SEM is only constant in symmetrical, mesokurtic score distributions (see Mollenkopf, 1949).

\(^{3}\) Ten years prior to this “discovery” made by Mollenkopf, Rulon (1939) showed that the SEM is equal to the standard deviation of the differences between the two half-test scores (which could be calculated at any score interval and thus resulting in the SEM at various levels). However, Rulon’s emphasis was not placed on the conditional nature of SEMs; the same was only a focus of the later work first conducted by Mollenkopf.
and IRT methods). Further, a study conducted by Feldt et al. (1985) shows that, while the difference methods are less complicated than some other methods (i.e., those based on ANOVA, IRT, or binomial methods), it is “integrated related” and generally produces very similar results.

IRT-based C-SEM methods typically require a large sample size for calibrating accurate parameter estimates (e.g., Tsutakawa & Johnson, 1990, recommend a minimum sample size of approximately 500 for accurate parameter estimates) and the use of sophisticated software (e.g., BILOG). Further, they can only be used for tests that meet certain psychometric assumptions (e.g., uni-dimensionality). The binomial error methods are generally less computationally intensive, and require only KR-21 reliability, the mean and standard deviation of test scores, and the number of items on the test. As “strong” methods, both of these techniques are tied to distributional assumptions.

The “difference” methods reviewed here focus on the Mollenkopf-Feldt (M-F) method, which uses polynomial regression to “smooth” the C-SEM values that are calculated using the variance between scores on two split test halves. Specifically, the M-F C-SEM method is calculated by first computing half-test scores for each examinee (e.g., using an odd-even split or some other method that breaks the test into two tau-equivalent halves) then using the total test score (X), total test score squared (X²), and total test score cubed (X³) to predict \( Y \) defined as:

\[
Y = \left( \frac{(X_1 - X_2) - (\bar{X}_1 - \bar{X}_2)}{X_1} \right)^3 \frac{1}{12}
\]

where \( X_1 \) and \( X_2 \) are the two test halves and \( \bar{X}_1 \) and \( \bar{X}_2 \) are the means for each test half. The resulting beta coefficients are then used to determine C-SEM values for each score in the distribution (the constant in the regression is set to 0 because this is the theoretical floor of C-SEM values).

Score bands can then be created by centering confidence intervals around classically computed Estimated True Scores, using the formula:

\[
ETS = (X-M) * r_{xx} + M
\]

(where \( X \) is the examinee’s obtained score, \( M \) is the average test score of all examinees, and \( r_{xx} \) is the reliability of the test) and the desired Confidence Interval (e.g., 1.96 for 95% confidence when considering 2 C-SEMs). This personnel score banding approach is based on the original M-F method and was modified for the personnel score banding (as originally defined in Biddle et al., 2007) and will be outlined in further detail in a forthcoming work by Biddle & Feldt (2008). Other members in this symposium have also contributed in many significant ways to the theoretical refinement and automation of the C-SEM banding methods presented in these works.

As a “weak” model, it is not tied to assumptions inherent to other models (e.g., IRT or the binomial methods). It has the added value of being tolerant of various item types (e.g., binary or polytomous items) and having sample size requirements similar to those of most multiple regression situations. One of the additional advantages of the M-F method is that it is based on the actual item responses and characteristics of each data set (binomial models can be computed without item-level data because they are based on the binomial probability distribution). The method allows data from each test administration to form varying C-SEM values at each score interval. Some research conducted by members on this panel shows that the M-F method is highly correlated (\( r = .88 \)) to the C-SEM models based on binomial methods.

The implications of using C-SEMs are wide and substantial for the I-O practitioner. Classical SEMs are widely used by personnel psychologists in two ways—
both of which have substantial impact on countless applicants that are ranked on various score lists: score banding and adjusting cutoff scores.

Test score banding is commonly practiced by I-O professionals as a way of preserving utility (over just using cutoffs) while minimizing adverse impact (over top-down ranking). Banding procedures typically use the Standard Error of Difference” (or “SED,” which is calculated by multiplying the SEM by the square root of 2), along with a confidence interval multiplier (e.g., 95% using a 1.96 multiplier) to categorize groups of “substantially equally qualified” applicants. The conditional nature of the SEMs, however, demands the use of C-SEMs when creating bands rather than using this traditional method.

Adjusting minimum competency cutoffs, as in the case of the modified Angoff technique, is another area where SEMs are commonly applied. In this situation, the predetermined minimum passing score for the test is lowered using 1, 2, or 3 SEMs to account for measurement error in the test (and thereby giving the “benefit of the doubt” to the applicant). In the experience of the panel members, most C-SEM techniques typically produce smaller values than classical SEM calculations, which translates to smaller groups of applicants being categorized because of the added precision of the conditional measurement precision (the classical SEM simply averages measurement error across the entire score range). Both of these will be discussed by the panel.
References


