5

Reliability in Educational Measurement

Won-Chan Lee
University of Iowa

Deborah J. Harris
University of Iowa (Retired)

Lee, W., & Harris, D. J. (2025). Reliability in educational measurement. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 277–381). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0005

This chapter is dedicated to exploring the concepts, quantification, and estimation of reliability and measurement errors, primarily within the framework of three prominent measurement theories: classical test theory (CTT), generalizability theory (GT), and item response theory (IRT). CTT, a well-established presence in the field of educational measurement, has consistently featured in previous editions of *Educational Measurement*. GT, having a shorter history than CTT, is often perceived as an extension or liberalization of CTT. The reliability chapters in prior two editions of *Educational Measurement* (Feldt & Brennan, 1989; Haertel, 2006) have provided relatively extensive treatments of GT. While the concept of reliability and measurement error can be discussed in the context of IRT, previous emphasis in IRT literature has typically centered around other aspects of the theory, such as model types, estimation of item and proficiency parameters, model fit, test construction, and scoring (see Hambleton, 1989; Yen & Fitzpatrick, 2006). With the growing applications of IRT in research and operational settings, it has become commonplace to present reliability information based on IRT models. Therefore, there is a pressing need to develop a cohesive framework for discussing reliability-related issues under the general assumptions of IRT. These three measurement theories offer distinct perspectives on various aspects of reliability, grounded fundamentally in their assumptions, which have significant consequences for the quantification and interpretation of reliability. This chapter primarily focuses on highlighting the similarities and differences among the theories in terms of conceptualizing and estimating various reliability statistics.

TERMS AND BACKGROUND

Measurement procedures are developed in accordance with the intended purposes of testing and the use of test scores for specific objects of measurement (e.g., test takers). This involves the identification of a set of tasks (e.g., multiple-choice items, essay prompts), administration modes (e.g., paper–pencil, computerized adaptive), scoring procedures (e.g., number correct, human rating), types of reported scores (e.g., summed raw scores, transformed scale scores), and score interpretations (e.g., norm-referenced, criterion-referenced). A specified set of measurement conditions constitutes a *test form*, with alternate forms encompassing different sets of similar (or parallel) measurement conditions. A sample of a test taker's responses to a test form is collected to generate one or more observed scores, which are then used to make general inferences about the test taker. The *observed score* for a test taker is considered a realization of a hypothetical distribution of all possible observed scores across repeated measurements using parallel forms of a measurement procedure. This observed score functions as an indicator or proxy for the underlying, unobservable *construct(s)* (i.e., domain of knowledge or skills) that a measurement procedure intends to measure. In IRT, proficiency estimates serve a similar role to observed scores.

Reliability, broadly conceived, is concerned with quantifying the extent to which observed scores are consistent over replications of a measurement procedure. However, the term *reliability* is often used to refer to *reliability coefficients*, which represent just one of many types of measures assessing the consistency of test scores. In this chapter, we make a clear distinction between these two terms. We use *reliability* to encompass a wide array of concepts and statistics used for quantifying the consistencies and/or inconsistencies in scores for individual test takers or groups of test takers. In a similar vein, the term *measurement error* is used broadly here to refer to both the sources of score inconsistencies and the summary statistics quantifying them for individual test takers or groups of test takers. It is evident that the characteristics of a measurement procedure play a crucial role in shaping the definition, interpretation, and estimation of statistics for reliability and measurement error. A particular focus will be placed on emphasizing the importance of reporting results for various statistics on the metric of *reported scores* used to make decisions about test takers. This is crucial because switching from one score scale to another, especially when the transformation is nonlinear, can lead to markedly different results and interpretations.

In his last paper, Cronbach (2004) stated,

> In the history of psychometric theory, there was virtually no attention to this distinction [between sample and population] prior to 1951. . . . It was not until Lord's (1955) explicit formulation of the idea of random parallel tests that we began to write generally about the sampling, not only of persons, but of items. This two-way sampling had no counterpart in the usual thinking of psychologists. No change in procedures was required, but writing had to become more careful to recognize the sample–population distinction. (pp. 401–402)

This statement recognizes Lord's significant contribution in bringing attention to the concept of *replication*, particularly in terms of sampling items. The distinction highlighted here is between an estimate of reliability specific to a particular sample of items and the desired population value that would be obtained over many other random sets of such items. The concept of replications becomes a central element in modern theory of reliability (Brennan, 2001a; Feldt & Brennan, 1989; Haertel, 2006), and it is used as a unifying framework throughout this chapter for discussing various theories, models, and estimators of reliability. The sample–population distinction mentioned in the above quote naturally draws attention to two overarching notions or phases in reliability analysis: *conceptualization* and *estimation*. Brennan (2001a, 2001b, 2006) similarly emphasized this distinction, albeit in a slightly different manner.

The notion of *conceptualization* refers to an investigator's conception of the intended use and interpretation of reliability, which calls for defining a set of measurement conditions for replications of the measurement procedure. The set of conditions allowed to vary across replications are the sources of inconsistencies in scores (i.e., measurement error), which need to be considered when estimating reliability. It is important to note that there is no universally correct or optimal definition of replications.

The determination of which sources of measurement error to include is at the discretion of the investigator and depends primarily on the intended uses and interpretations of test scores.

The *estimation* phase involves quantifying reliability estimates (i.e., estimated reliability statistics) that best capture the characteristics defined in the conceptualization phase and requires decisions on the appropriate reliability statistics and data collection designs. The meaningfulness of interpretations from a reliability estimate hinges on how faithfully the specifications in the conceptualization phase are reflected in the estimation phase. Thus, it is essential to choose and compute a reliability estimator using an appropriate data collection design that effectively incorporates, within practical constraints, the influence of the specified sources of measurement error deemed important by the investigator. If the chosen reliability statistic involves sources of measurement error different from those conceptualized, an explanation is necessary. This is because the results could either overestimate or underestimate the parameter of interest, and understanding any divergence is crucial for accurate interpretation.

Consider, for example, an investigator wanting to estimate how consistent the rank orders of IRT proficiency estimates for candidates applying to a degree program remain when tested with a different set of multiple-choice items, possibly at different times. The measurement procedure involves the target population, exam characteristics, score metric, and score interpretation. The conceptualization envisions replication involving the use of different items and testing occasions.

For a direct estimation, data from two alternate test forms are collected from the same group at two separate occasions, applying the same IRT scoring method to both data sets. The estimated reliability coefficient is derived from the correlation between the proficiency estimates, aligning with the investigator's conceptualization of using different items and occasions.

Now, suppose the investigator chooses to report coefficient alpha based on a single administration with number-correct scores. This fails to address the original question, given that the metric of interest is IRT proficiency, not number-correct scores. Additionally, coefficient alpha overlooks testing occasions as a source of error, likely resulting in an underestimate of error variance and, consequently, an overestimate of the reliability coefficient. This example underscores the importance of selecting a reliability estimator that aligns with the conceptualized replications and is consistent with the intended use of the measurement procedure.

The preceding discussion on the conceptualization-estimation scheme is mostly aptly characterized within the framework of GT. However, the same principle is applicable to reliability analyses under CTT and IRT. The key distinction lies in the fact that, in CTT and IRT, investigators often select a reliability estimator and/or a data collection design without giving serious consideration to what constitutes replications. In effect, the investigator is making assumptions about replications, whether intentionally or unintentionally, through their choices. It should also be noted that the three measurement models differ not only in their mathematical representations, but also in the framework for conceptualizing replications, true score, and measurement error, leading to potential ambiguities when interpreting results. The following section discusses some of the conceptual similarities and differences among the three measurement models regarding the definitions of true score and measurement error.

Throughout this chapter, the term *consistency* is primarily used to depict the variability of scores over replications. The term *precision* has also found use in the literature, including in the current *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association [AERA] et al., 2014), particularly in the context of reliability. Although the term precision literally denotes *exactness*, which diverges in meaning from the concept of consistency, it is occasionally used in this chapter when the context deems it more suitable. The term *accuracy* is specifically reserved for one of the classification indices, known as classification accuracy. This usage is retained to describe the degree of *correctness* of a measurement relative to the true value.

NOTIONS OF TRUE SCORE AND MEASUREMENT ERROR

Comprehending the concept of reliability requires grasping the notions of true score and measurement error. Almost all measurements in scientific disciplines involve observing the objects of measurement under certain conditions to produce *observed scores*. When observed under different conditions, the resulting observed scores are likely to differ, even if both measurements are intended to measure the same construct on the same objects of measurement. *True score*, by contrast, cannot be directly observed; rather, it is *defined*. True scores are pivotal in reliability, and how they are defined significantly affects the estimation and interpretation of reliability statistics. CTT and GT share similar perspectives on true score, while IRT employs a few distinct versions of true score.

In essence, true score in both CTT and GT is perceived as the *expected value* of observed scores over replications of a measurement procedure. This perspective precludes the platonic interpretation, which considers true score as a definitive indication of what the measurement intends to assess. The platonic view of true score has faced criticism because of its limited scientific utility, particularly in the context of psychological and educational measurement, where the constructs of interest are often intricate and challenging to explicitly define (see Lord & Novick, 1968, chap. 2).

The expected-value notion of true score requires specifying what constitutes a replication, implying that there is no singular definition of true score. The investigator has control over defining parallel forms for replications, such as different sets of items, raters, and/or occasions. Many well-known results in CTT rely on the assumption that replications are performed over forms that are *classically parallel*.

In contrast, later developments of single-administration reliability coefficients use less stringent assumptions, such as essentially tau-equivalent or congeneric forms. Meanwhile, true score in GT (also referred to as universe score) is defined as the expected value of observed scores over measurement conditions or forms that are *randomly parallel*. This distinct conception of parallelism for replications stands out as a key difference between the two theories, carrying considerable theoretical and practical implications.

The term *measurement error* refers to the discrepancy between a test taker's true score (i.e., the average of observed scores over replications) and the actual observed score. This implies that the definition of measurement error depends on how true score is conceptualized. Measurement error arises due to the random variation of observed scores over repeated testing using specified conditions of a measurement, which are used to define true score. Other error sources, deemed less important and thus not explicitly modeled by the investigator, are assumed to operate randomly. In CTT and GT, it is assumed that specified and unspecified (i.e., residual) sources of measurement error behave randomly such that their expected value over replications is always zero. If there are factors unrelated to the construct the test aims to measure that influence test-taker scores in some *systematic* way (e.g., verbal skills in math computation, bias in raters, model misfit), they are considered sources of construct-irrelevant variance and are deemed threats to *validity* (Kane, 2006). Rarely does construct-irrelevant variance play any role in estimating reliability under CTT or GT. However, it does hold implications in the context of IRT, as discussed next.

In unidimensional IRT, the notion of true score is conceptually linked to the person proficiency parameter, denoted θ. Although the proficiency parameter is explicitly defined and specific to a particular IRT model, subsequent discussions are intended to be general, applicable to any model. The proficiency parameter θ represents a person's location in a latent trait space and plays a role similar to that of true score. Consider a biased estimator of θ, such as the maximum likelihood estimator. Due to this bias, the expected value of the estimates across replications, denoted $\mathscr{E}\hat{\theta}$, does not equal the parameter θ —that is, $\mathscr{E}\hat{\theta} \neq \theta$. Unlike CTT, which adopts the expected-value notion of true score, the two possible definitions of a person's proficiency in IRT (i.e., θ and $\mathscr{E}\hat{\theta}$) contribute to some inconsistencies and ambiguities in estimating and interpreting reliability coefficients for IRT proficiency estimates. In this chapter, θ is referred to as *latent proficiency*, while $\mathscr{E}\hat{\theta}$ denotes *expected proficiency*, the latter being considered a more suitable definition in the context of reliability. Clear distinctions between these two definitions are not always made, and both have been used in various formulas for estimating reliability in IRT.

Different definitions of true proficiency inevitably result in distinct definitions of measurement error. Error of measurement, defined by the discrepancy between $\hat{\theta}$ and $\mathscr{E}\hat{\theta}$, constitutes random error, and its variance is referred to here as the *expected error variance*. By contrast, the variance of discrepancies between $\hat{\theta}$ and θ—referred to simply as *error variance* here—involves both systematic error (i.e., bias) and random error. While bias can diminish validity, it does not impact reliability. If the variance (and covariance) of bias is incorporated into a reliability coefficient, it should contribute to true score variance rather than error variance, thereby increasing reliability. In practice, there is often an argument that the bias in proficiency estimates, particularly its variance, is negligible, making the distinction unnecessary. However, empirical examples presented by W. Lee et al. (2025) demonstrate that differences can sometimes be substantial.

In IRT, the definition of true score expressed in the number-correct (or summed raw) score metric is represented by a test characteristic curve (TCC). The TCC is a nonlinear transformation of θ, where values of θ with an unbounded range from $-$ to ∞ are transformed into a score range of 0 to the maximum possible summed raw score. A TCC value represents the expected value of the model-based distribution of observed scores for a given θ, assuming known item parameters. Measurement error is characterized by the variability of model-based observed scores for a given θ. The TCC definition of true score in IRT is akin to true score in CTT in two key aspects: (a) It is expressed on the summed raw score metric, and (b) it employs the expected-value notion of true score as opposed to the latent-trait definition. Consequently, most results derived in CTT remain applicable for estimating reliability on the metric of summed raw scores in IRT. It is crucial to note that, similar to $\hat{\theta}$ (and θ), a TCC is model-specific, signifying that it is meaningful and can be interpreted properly only under the chosen IRT model.

The consideration that a TCC for a given model is often deemed fixed implies that the expectation is taken over forms containing a set of items with identical item parameters. This argument also extends to the expected proficiency, $\mathscr{E}\hat{\theta}$. Forms with identical item parameters are termed *strictly parallel* forms, effectively implying a fixed form. As a result, reliability coefficients under IRT should be larger than those based on CTT or GT, all other factors being equal.

TYPES OF SCORES CONSIDERED

Scores derived from test-taker responses can manifest in various forms, and the choice of a reporting metric introduces distinct considerations for reliability analyses. In this chapter, our focus centers on specific types of scores:

- *Summed raw scores:* These scores result from the summation of points earned on each individual test item. Variants include percent correct scores and weighted sum scores, where items may be assigned different weights based on content importance or other considerations. Summed raw scores may be referred to simply as *raw scores* in this chapter.
- *IRT proficiency estimates:* Derived from IRT models, proficiency estimates are based on test takers' responses and item parameters. While profi-

ciency estimates are not commonly used for reporting because of their interpretational complexity, they find frequent application in adaptive testing to guide item selection. Two different IRT proficiency estimators are considered in this chapter based on the maximum likelihood and Bayesian estimation methods.

- *Composite scores:* Composite scores are linearly weighted sums of component scores, either subscores from one test or scores from different tests. Multiple components can be characterized by the differences in content areas, constructs measured, item formats, and other aspects. Unlike simple summed raw scores, composite scores treat items across components distinctly.
- *Scale scores:* Widely employed for reporting, scale scores can be crafted in various ways, such as setting target means and standard deviations, incorporating precision information for equal conditional standard errors of measurement, or setting a cut score with a particular scale score value. Scale scores prove particularly useful in reporting when test takers take different sets of items. A conversion table is typically created to convert summed raw scores, IRT proficiency estimates, or composite scores to scale scores.
- *Classification category scores:* Instead of providing a specific numerical score, classification category scores convey a test taker's performance category, such as *Exceeds Expectations, Meets Expectations*, or *Below Expectations*. *Pass/Fail* or *Master/Nonmaster* are also common in various assessments.

STATISTICS AND INDICES FOR RELIABILITY

Multiple methods exist for quantifying reliability for individual scores or scores for a group of test takers, and some of these are defined next.

Overall Standard Error of Measurement

The overall standard error of measurement (SEM) represents the standard deviation, over individuals, of observed scores minus true scores for an assessment. Traditionally, it is estimated as a function of a reliability coefficient and observed score variance. This estimation is referred to as the overall SEM, distinguishing it from the conditional SEM. SEMs, both overall and conditional, are expressed in the same units as the reported score, making them specific to the scoring metric used. Consequently, SEMs cannot be directly compared across different scoring procedures. Another notable characteristic of SEM is its relative insensitivity to the characteristics of a specific group of test takers.

Conditional Standard Errors of Measurement

The SEM typically varies across score levels or test takers. The conditional SEM (CSEM), in theory, represents the standard deviation of observed scores over repeated measurements conditional on a test taker's true score. In practice, estimated CSEMs can be reported either at each true score level based on a psychometric model or for each individual test taker using an observed score as an estimate of the true score. CSEMs provide valuable information about the amount of measurement error for each test taker (or score level) and can also be aggregated over all score levels to compute the overall SEM and reliability coefficients. The section "Estimators of CSEMs" of this chapter is dedicated to discussions on estimators of CSEMs for various types of scores.

Reliability Coefficients

The extent to which test takers' scores are consistent over replications can be quantified using reliability coefficients. These coefficients, originally developed under the traditional assumptions of CTT, take various forms. One such coefficient, for example, involves the correlation between observed scores on the same assessment or parallel forms of an assessment. Reliability coefficients are not reported in score units, making them challenging for users to interpret directly. Reliability coefficients are sensitive to the characteristics of the test-taker group. Some estimators require at least two scores per test taker, which is a challenging requirement to meet in practice, while others can be computed based on a single administration of an assessment. Various approaches to estimating reliability coefficients under each of the three measurement model frameworks are discussed in this chapter.

Classification Consistency and Accuracy Indices

When test scores are used to categorize test takers based on one or more cut scores, a crucial consideration is the likelihood of consistent classification if the test is administered again. Classification consistency serves as a criterion-referenced measure of reliability, assessing whether test-taker performance aligns with established standards or cut scores across replications. In contrast, classification accuracy is concerned with whether a test taker is "accurately" categorized into the performance category corresponding to their true score. This definition implies a closer connection to validity. Because consistency and accuracy are often considered together, both aspects are addressed in this chapter. The section "Reliability of Classification Category Scores" of this chapter is dedicated to presenting methods for estimating various classification consistency and accuracy indices.

ORGANIZATION OF THIS CHAPTER

This chapter builds on concepts discussed in earlier editions of *Educational Measurement*, particularly in Haertel (2006) and Feldt and Brennan (1989). The sections "Reliability in CTT" and "Reliability in GT" delve into classical and generalizability theory approaches. The section "Reliability in IRT" treats IRT approaches to reliability issues, offering a more extensive discussion because of the relatively less explored nature of this area. Some content discussed in the IRT section is novel and has not been previously published. Emphasizing the utility of CSEMs over reliability coefficients, the section "Estimators of CSEMs" provides various estimation procedures for CSEMs. The section "Reliability of Classification Category Scores" focuses on reliability of classifications, while the section "Other Models, Aggregation, and Precision Issues" tackles various additional issues associated with reliability. The final section offers a concise summary of the chapter and outlines potential areas for future research on reliability. For practical applications and comparisons of different methodologies, readers can refer to two real data examples presented by W. Lee et al. (2025), which demonstrate the computation of various reliability and error statistics across different models and metrics.

RELIABILITY IN CTT

CTT emerged in the early 20th century, initially focusing on studies of individual differences. Its origins are often traced back to Spearman's (1904) work on methods for correcting correlation coefficients for attenuation due to measurement error (Traub, 1997). Since then, numerous scholars have contributed to further developing and refining the theory. Key texts in the field, such as Gulliksen (1950) and Lord and Novick (1968), are considered highly influential, offering comprehensive treatments of CTT.

Despite its relatively simple set of assumptions and definitions compared to other contemporary theories, CTT yields results that are relevant to many aspects of modern psychometric applications and remains widely used. Not surprisingly, CTT continues to be an important area of study. In recent decades, certain segments of the measurement literature have extended CTT substantially, particularly in areas such as estimating CSEMs for both raw and scale scores, classification consistency and accuracy, and interval estimation (although the latter is not extensively covered in this chapter).

Assumptions, Definitions, and Basic Results

CTT asserts that an observed score for person p on form f, denoted X_{pf}, can be decomposed into two components: a true-score component, T_p, and an error-score component, E_{pf}:

$$X_{pf} = T_p + E_{pf}, \tag{1}$$

where the true score T_p is a constant specific to the person and does not depend on forms. For notational convenience, T_p and τ_p are used interchangeably in this chapter to represent the true score for person p in the total raw-score metric. The intrinsic difficulty of using the model in Equation 1 arises from the fact that only one variable can be observed, while the other two are unobservable. This challenge is circumvented by making certain assumptions about the unobservable true or error scores—assuming either one of them to be known makes the other apparent. One such assumption is that, for a specific person p, the expected value ($\mathscr{E}$) of error scores over replications is zero: $\mathscr{E}_f(E_{pf}) = 0$, where the expected value is taken over a hypothetically infinite number of parallel forms, as indicated by the subscript *f* below the expectation operator. Under this assumption, it follows that

$$\mathscr{E}_f(X_{pf}) = \mathscr{E}_f(T_p + E_{pf}) = T_p. \tag{2}$$

Conversely, if the true score is defined as the expected value of observed scores, the expected value of errors is necessarily zero. Therefore, the true score in CTT, which is equal to the expected value of observed scores, can be viewed either as a derived result or as a definition.

Assuming the measurement procedure does not alter a person's true score, any variation in the observed scores over replications is attributed solely to the use of parallel, yet different forms of a test. The question of what constitutes the replications over which the expectations in Equation 2 are taken lies at the core of CTT because the nature of true score and measurement error depends on the answer to that question. One traditional answer, among many possibilities, is that expectations are taken over forms of a test that are *classically parallel*. However, as discussed later, different definitions of replications can lead to different results. It should be noted that there is no universally right or best definition of replications, and the choice depends entirely on the investigator's decision, guided by the intended interpretations of test scores.

Another assumption about error scores is $\mathscr{E}_p(E_{pf}) = 0$, meaning that, for a given test form, the expected value of the errors taken over a population of persons equals zero. This assumption holds for any subpopulation of persons, unless persons are selected based on the magnitude of X_{pf}. Under this and other assumptions discussed previously, the following central results can be derived: (a) true scores and errors for any test form are uncorrelated; and (b) error scores on any pair of parallel forms, *f* and *g*, are uncorrelated. It is also true that their covariances are zero, notationally, $\sigma(T, E_f) = 0$ and $\sigma(E_f, E_g) = 0$. It follows that the variance (over persons) of observed scores on a form is simply the sum of true score variance and error score variance:

$$\sigma^2(X_f) = \sigma^2(T) + \sigma^2(E_f). \tag{3}$$

It can be further shown that, under CTT assumptions, true score variance is equal to the covariance between observed scores on a pair of forms and also to the covariance between observed scores and true scores: $\sigma(X_f, X_g) = \sigma(X_f, T) = \sigma^2(T)$.

Reliability Coefficients and SEM

The historical definition of the reliability coefficient is rooted in the conceptually straightforward notion of replication, specifically the correlation between two parallel measurements. The notation $\rho(X_f, X_g) \equiv \rho(X, X')$ denotes the reliability coefficient, defined as the correlation between observed scores arising from the same form or two parallel forms. Under the CTT assumptions and results set forth previously, other expressions of the reliability coefficient can be derived (note that, without loss of generality, the subscript *f* for test form will be dropped hereafter):

$$\rho(X,X') = \rho^2(X,T) = \frac{\sigma^2(T)}{\sigma^2(X)} = 1 - \frac{\sigma^2(E)}{\sigma^2(X)}. \tag{4}$$

It is important to note that these equalities are derived under specific assumptions in CTT, particularly those pertaining to classically parallel forms, and should not be universally applied without consideration. For instance, as detailed later, alternative expressions of the reliability coefficient may produce different results in certain IRT contexts. Additionally, in this chapter, the terms *definitions* and *expressions* are used interchangeably to offer clearer distinction in various situations.

Using $R(X)$ as a generic notation for the reliability coefficient in the raw-score metric X, the following symbols distinguish various expressions presented in Equation (4):

$$R(X)_{PF} \equiv \rho(X, X')$$

$$R(X)_{SC} \equiv \rho^2(X,T)$$

$$R(X)_{VR} \equiv \sigma^2(T)/\sigma^2(X)$$

$$R(X)_{EV} \equiv 1 - \left[\sigma^2(E)/\sigma^2(X)\right]$$

This distinction gains increased importance and utility in IRT contexts. The parallel-forms definition, $R(X)_{PF} \equiv \rho(X,X')$, stands apart from the others by being expressed solely in terms of observable quantities without referencing the true or error component. One limitation of this definition is the absence of an explicit parameter for the reliability coefficient because the correlation may vary across different pairs of forms unless forms are definitely classically parallel.

In the extant literature, the squared-correlation definition, $R(X)_{SC} \equiv \rho^2(X,T)$, is regarded as the canonical definition of the reliability coefficient (Brennan, 2010). The generalizability coefficient in GT, which is analogous to the reliability coefficient in CTT, takes the form of the squared correlation. The variance-ratio definition, $R(X)_{VR} \equiv \sigma^2(T)/\sigma^2(X)$, employs the ratio of true score variance to observed score variance, indicating that if the quantity is close to 1.0, variability in test takers' observed scores is largely attributable to variability in their true scores rather than errors. This definition forms the basis for the development of many internal consistency reliability coefficients (e.g., coefficient alpha), where a key objective is to express unobservable true score variance in terms of some observable quantities. The last definition involving error variance, $R(X)_{EV} \equiv 1 - \left[\sigma^2(E)/\sigma^2(X)\right]$, always yields the same result as the variance-ratio definition, as long as Equation 3 holds. This definition may be considered the most general form of the reliability coefficient because of its broad applicability to various score types.

The reliability coefficient essentially represents a correlation ranging from zero to 1, which is a convenient unitless scale for comparing different instruments or measurement procedures. By contrast, SEM is expressed in units of a reported score scale and describes the extent to which test takers' scores vary from one testing to another. Traditionally, SEM, in the raw-score metric, is derived through the reliability coefficient as

$$\sigma(E) = \sigma(X)\sqrt{1 - R(X)}. \tag{5}$$

Strictly speaking, this derivation holds only if the assumptions of classically parallel forms are met. However, in practice, these assumptions are often disregarded and loosely applied in conjunction with various reliability coefficient estimates. The consequence of violating these assumptions is relatively unknown.

From the perspective of CTT, the SEM derived in Equation 5 is termed the *overall* SEM for a population of persons. Alternatively, the overall SEM can be defined as the square root of the expected value of individual-level error variances, which typically differ across persons with different true score levels. Symbolically,

$$\sigma(E) = \sqrt{\mathscr{E}_p \sigma^2(E \mid \tau_p)} = \sqrt{\mathscr{E}_p \sigma^2(X \mid \tau_p)}, \tag{6}$$

where $\sigma^2(E \mid \tau_p) = \sigma^2(X \mid \tau_p)$ is called the *conditional* error variance for persons with fixed true score τ_p, and the square root of it is the CSEM. Numerous theoretical and empirical studies have highlighted that CSEMs exhibit variation along the score scale, raising concerns about the applicability of the overall SEM to all individuals. (Refer to the section "Estimators of CSEMs" for estimators of CSEMs for various score types under all three measurement models.) It is important to note that the exact mathematical relationship between the overall and conditional SEMs in Equation 6 holds only for certain models and estimators.

In the measurement literature, reliability coefficients have been the central focus, with insufficient attention given to the role of CSEMs. CSEMs not only provide information about the amount of measurement error specific to each individual or score point but also can be used to estimate overall statistics such as the overall SEM and reliability coefficients. They can further be integrated into interval estimation procedures (K. Y. Kim & Lee, 2018; W. Lee et al., 2006). In addition, obtaining CSEMs for scale scores is relatively straightforward.

Approaches to Estimating Reliability

Approaches to estimating the reliability coefficient and SEM in CTT can be broadly categorized into two types. In the first type, reliability is directly estimated using data from two independent replications of the "full-length" measurement procedure administered to a group of test takers. The resulting two sets of observed scores at the group or individual levels are used for computing reliability coefficients, SEM, CSEMs, and other relevant statistics. In contrast, the second type estimates reliability using data collected from a single test administration to avoid the challenges of conducting two actual replications. Most estimators of the second type involve splitting a full-length test form into two or more constituent part-tests. The data from these part-tests from a single group of test takers are treated as replications—a somewhat contrived notion of replications. Statistical quantities derived from part-tests are then used to obtain results for the full-length test form. However, this approach relies on a strong set of assumptions regarding the parallelism of the part-tests.

In principle, data for a reliability analysis are collected in a way that properly reflects the influences of all important sources of error in the reliability statistics of interest. To generalize interpretations of test results beyond a particular incident of a measurement procedure, investigators would need to conceptualize which measurement conditions shall vary over replications and design data collection accordingly. Ideally, in the data collection process, only the measurement conditions contributing to the conceptualized measurement error should be allowed to vary over replications, while all other potential sources of random error remain unchanged or are treated as negligible.

Direct Estimation With Full-Length Replications

Suppose an investigator is interested in generalizing test takers' test scores beyond a particular collection of items on a test form. A straightforward approach is to construct two full-length parallel forms of the test according to the same test specifications. These two forms are then administered (possibly in a counterbalanced manner) to the same group of test takers within a very short time interval, such as on the same day. The correlation between two sets of scores is subsequently calculated. This correlation value serves as a direct estimate of the reliability coefficient for either test form (not the sum or average of the two). Known as a *coefficient of equivalence*, this correlation indicates the extent to which test takers perform similarly on different forms of the test.

When two test forms are administered at different times (e.g., a few days apart), the resulting correlation is affected downward due to the measurement error arising from differences in both testing occasions and test forms. This correlation is referred to as a *coefficient of stability and equivalence*, which is often preferred over other estimates because investigators typically seek to generalize results over different testing occasions as well as different test forms. In this case, the coefficient of equivalence might be an overestimate of the idealized (or conceptualized) reliability. Alternatively, in situations where only one test form is available and administered twice, typically at different times, the estimate is called a *coefficient of stability* or *test–retest reliability coefficient*. This coefficient reflects variability in observed scores over testing occasions, which may be biased by memory, but it does not account for variability due to different test forms. It is important to emphasize that these coefficients are not interchangeable principally because they involve the impact of different sources of measurement error. The choice of which coefficient to use depends on its consistency with the intended generalization.

The direct estimation approaches using actual full-length replications offer several advantages. First, in theory, they can be applied to any measurement procedure, including performance assessments involving raters, as long as resources permit. Second, these approaches require minimal statistical and psychometric assumptions, although they do rely on other assumptions such as no change in true score or latent traits between administrations, no memory effects on the second responses, and no fatigue or practice effects. Third, the computation is easy and straightforward. Fourth, other reliability statistics, including CSEMs, can be readily computed from the replicated data. Fifth, it provides the investigator an opportunity to carefully consider what constitutes replications of the measurement procedure, with a central focus on the sources of measurement error (e.g., items, occasions, raters). Despite all these benefits, however, direct estimates are less frequently used in practice because of their requirement for double testing time and, for some coefficients, the construction of an additional parallel form.

Estimation Based on a Single Test Administration

Spearman (1910) and Brown (1910) were the first to recognize the necessity of estimating a reliability coefficient using data from a single test administration. Since then, numerous estimators have been developed, collectively known as *internal consistency coefficients*. The well-known Spearman–Brown formula was developed for the purpose of estimating a reliability coefficient based on two split-halves with equal observed score means and variances (i.e., classically parallel). Recognizing the challenge of constructing part-tests with precisely identical means and variances, subsequent developments attempted to relax the strict assumptions of classically parallel forms. In particular, Feldt and Brennan (1989) integrated various internal consistency coefficients and categorized them based on varying degrees of part-test parallelism.

The notions of tau-equivalent and essentially tau-equivalent forms were initially introduced by Lord and Novick (1968). For two forms (or part-tests) that are *tau equivalent*, it is assumed that true scores are the same, thereby necessitating equal observed score means. However, these two forms are allowed to have different error variances, leading to the possibility of differing observed score variances. Under *essentially tau-equivalent* forms assumptions, the requirement of equal observed score means can be relaxed further. Here, true scores may differ by a constant, allowing for variations in observed score means. Despite these differences, true score variances are equal for essentially tau-equivalent forms due to the constant difference in true scores. A less stringent concept of parallelism is *congeneric* forms, where true scores are assumed to have a linear relationship. Consequently, true score variances may vary, but they are perfectly correlated. Under the congeneric assumptions, observed score variances may also differ due to variations in both error and true score variances.

Extensive discussions and derivations of internal consistency coefficients, considering various definitions of form parallelism, can be found in Feldt and Brennan (1989) and Haertel (2006). In this section, a subset of these coefficients will be briefly introduced, starting with those based on two parts and then extending to coefficients involving more than two parts.

For two-part coefficients, the full-length test X is divided into two scorable part-tests, X_1 and X_2, such that $X = X_1 + X_2$. The Spearman–Brown formula, the first internal consistency coefficient in the history of measurement, was derived under the assumption that X_1 and X_2 are classically parallel. The correlation between the two part-tests, denoted by $\rho(X_1, X_2)$, serves as a reliability coefficient for a half-test. This coefficient is then stepped up to obtain a reliability coefficient for the full-length test X using the simple formula

$$ {}_{SB}R(X) = \frac{2\rho(X_1, X_2)}{1 + \rho(X_1, X_2)}. \tag{7}$$

Other two-part coefficients developed subsequently relax the classically parallel form assumptions to express the unobservable true score variance in terms of observable quantities based on two parts. Rulon (1939) provided a formula (attributed to Flanagan) under the assumption of essential tau equivalence, where the true score variance for the total test is expressed to be equal to four times the covariance between X_1 and X_2: $\sigma^2(T_X) = 4\sigma(X_1, X_2)$. Algebraically identical versions were proposed by Rulon (1939) and Guttman (1945). Putting all three versions together, the Flanagan–Guttman–Rulon coefficient is

$$ {}_{FGR}R(X) = \frac{4\sigma(X_1, X_2)}{\sigma^2(X)} = 2\left(1 - \frac{\sigma^2(X_1) + \sigma^2(X_2)}{\sigma^2(X)}\right) = 1 - \frac{\sigma^2(X_1 - X_2)}{\sigma^2(X)}. \tag{8}$$

The assumption of equal true score variance in essential tau equivalence is rarely satisfied when, for example, two parts have substantially different test lengths or are associated with different content areas. Allowing for different part-test true score variances increases the number of unknown quantities and calls for additional constraints to find a solution. Raju (1970) introduced a constraint that the relative lengths of the two parts, l_1 and $l_2 = 1 - l_1$, are known (e.g., the actual number of items in each part-test). In contrast, Angoff (1953) and Feldt (1975) independently proposed the same solution for situations in which the lengths of part-tests are treated as unknown parameters to be estimated. They used a constraint that the part-test error variances be linearly related to each other as a function of the l terms. It follows that the unknown part-test lengths are equal to *effective test lengths*: $l_1 = \left[\sigma^2(X_1) + \sigma(X_1, X_2)\right] / \sigma^2(X)$ and $l_2 = \left[\sigma^2(X_2) + \sigma(X_1, X_2)\right] / \sigma^2(X)$. The Angoff–Feldt coefficient is given by

$$_{AF}R(X) = \frac{\sigma(X_1,X_2)}{l_1 l_2 \sigma^2(X)} = \frac{4\sigma(X_1,X_2)}{\sigma^2(X) - \left(\frac{\sigma^2(X_1) - \sigma^2(X_2)}{\sigma(X)}\right)^2}. \tag{9}$$

Note that Raju's coefficient is a special case of Equation 9 where l_1 and l_2 are predefined constants.

Among many possible methods for splitting a test into two scorable parts, the odd–even split is most frequently employed in practice. The goal in dividing a test is to create two half-tests, each of which reflects the specifications of the full-length test and is as similar as possible to each other in content and item characteristics (e.g., difficulty, format) to closely adhere to parallelism assumptions. Failure to achieve this similarity may introduce bias into estimates of reliability. An inherent limitation of the two-part approaches is that, regardless of the split-half strategy used, there are many possible split halves that are equally acceptable. However, not all such splits will produce the same estimate.

Subsequently, the two-part approaches were extended to multiparts. Multipart estimates, often derived from individual items treated as parts, are generally preferred over two-part estimates because they avoid arbitrary divisions of a test, and the sampling errors in reliability coefficients tend to be smaller (Kristof, 1963). However, two-part approaches may be more defensible than multipart approaches concerning parallel-part assumptions. For example, the claim that all items are essentially tau equivalent is rarely justifiable because it implies that any differences in the item-level observed score variances are solely due to measurement error. Constructing two parts that largely satisfy essential tau equivalence might be much easier than developing all individual items with such a high degree of parallelism. Therefore, multipart procedures should not necessarily be perceived as "better" than two-part procedures.

Suppose a full-length test X can be divided into n parallel part-tests: $X = X_1 + X_2 + \cdots + X_n$, where each part-test can be an individual item or a cluster of items. Under the assumption of classically parallel part-tests, the generalized Spearman–Brown (GSB) formula can be derived as

$$_{GSB}R(X) = \frac{nR(X_1)}{1 + (n-1)R(X_1)}, \tag{10}$$

where $R(X_1)$ is a reliability coefficient for a unit-length test (i.e., a part-test). If the above equation is solved for $R(X_1)$, one can obtain a prophecy formula for predicting the reliability of a shortened test. Therefore, a more general formula can be written as

$$_{pred}R(X) = \frac{\upsilon \times {}_{orig}R(X)}{1 + (\upsilon - 1) \times {}_{orig}R(X)}, \tag{11}$$

where $_{orig}R(X)$ is the reliability coefficient of the original test, and υ is the ratio of the length of the predicted test to the length of the original test. Obviously, υ can be any real number greater than zero.

Cronbach's (1951) alpha is undeniably one of the most widely recognized multipart reliability coefficients, assuming essentially tau-equivalent part-tests.[1] The formula is given by

$$_{\alpha}R(X) = \left(\frac{n}{n-1}\right)\frac{\sigma^2(X) - \sum\sigma^2(X_i)}{\sigma^2(X)}, \tag{12}$$

where $\sigma^2(X_i)$ is the variance of each part-test, $X_i (i = 1,\ldots,n)$. Note that the multiplier $n/(n-1)$ has nothing to do with correcting for bias. Notably, coefficient alpha is algebraically equal to Kuder and Richardson's (1937) KR20 coefficient if each item is treated as a part-test, and all items are scored dichotomously (i.e., 0 or 1). Kuder and Richardson's KR21 is a special case of KR20 when items are all equally difficult. However, because of the unrealistic assumption of equal item difficulty, the practical utility of KR21 is limited.[2] Another well-known property of $_{\alpha}R(X)$ is that it equals the average of all possible split-half reliability coefficients computed using $_{FGR}R(X)$ in Equation 8. As a special case, when $n = 2$ in Equation 12, $_{\alpha}R(X) = {}_{FGR}R(X)$.

Reporting coefficient alpha is prevalent and nearly a standard practice in the measurement and assessment literature. However, it is also true that alpha is frequently misused and/or misunderstood. One example of improper use of coefficient alpha is to assess the dimensionality of items in a test. Nothing in the derivation of alpha requires unidimensionality—alpha is not based on a latent trait model. In his reflection on coefficient alpha 50 years later, Cronbach (2004) stated, "I particularly cleared the air by getting rid of the assumption that the items of a test were unidimensional" (p. 397).

Another widely cited but potentially misleading characteristic of alpha is the phrase that alpha is a lower limit to reliability (Lord & Novick, 1968). While the mathematical proof is valid under a particular set of assumptions, it is hardly generalizable to other measurement circumstances for which the sources of error are different from those involved in alpha. For example, if the investigator intends to generalize over different testing occasions in addition to different forms, coefficient alpha computed based on data from a single occasion will likely be an overestimate (Brennan, 2001b, 2010). As mentioned earlier, the crux of the matter is whether the data used for estimation reflect the effects of error sources to be consistent with the investigator's conceptualization of replications. Cronbach (2004) also stated, "I no longer regard the alpha formula as the most appropriate way to examine most data" (p. 403). In theory, depending on the investigator's conceptualization, there can be many reliability coefficients for any set of test scores, and Cronbach incorporated this notion into the development of GT.

Another coefficient for essentially tau-equivalent parts, developed by Guttman (1945), is coefficient λ_2, which is one of Guttman's series of lower bounds for reliability referred to as λ_1 through λ_6. Guttman's λ_2 is always greater than or equal to λ_3, which is the same quantity as alpha, but is lower than "the" reliability. In this sense, λ_2 is often considered a "better" lower bound than alpha. However, again, the conceptual definition of reliability with respect to sources of error for replications should take precedence over any statistical justification for lower bounds.

If the stringent assumptions of essentially tau-equivalent parts are relaxed to be congeneric, part-tests are allowed to have heterogeneous error variances and true score variances, while maintaining a perfect correlation between true scores. The congeneric model is deemed suitable for cases in which observed part-test variances differ considerably, and one cannot solely attribute such variations to measurement error. For example, consider a test comprising several units (e.g., reading passages), each associated with a different number of items. If these units serve as part-tests in a reliability analysis, the observed part-test variances may well differ, and the differences are likely attributable not only to measurement error but also to variations in true score variances.

Under the congeneric assumptions, Raju (1977) extended his two-part coefficient (Raju, 1970) to multipart contexts, assuming the proportions of total test length for part-tests, l_i, are known:

$$_{R}R(X) = \left[\frac{1}{1 - \sum l_i^2}\right]\left[\frac{\sigma^2(X) - \sum\sigma^2(X_i)}{\sigma^2(X)}\right] \tag{13}$$

which becomes identical to coefficient alpha if $l_i = 1/n$. For unknown part-test lengths, Feldt and Brennan (1989) provided a solution called Feldt's coefficient:

$$_{F}R(X) = \frac{\sigma^2(X)\left[\sigma^2(X) - \sum\sigma^2(X_i)\right]}{\left[\sigma^2(X)\right]^2 - \sum[\sigma(X_i, X)]^2}. \tag{14}$$

The derivation of Feldt's coefficient requires congeneric part-tests and an assumption that error variances follow dictates of CTT, meaning the error variance for each part-test equals l_i times the error variance for the total test—this model is often referred to as *classical congeneric*. The formula for Feldt's coefficient can also be obtained by using the effective test lengths, $l_i = \sigma(X_i, X)/\sigma^2(X)$, in Equation 13.

Other congeneric-model coefficients that are not discussed here in detail include Kristof's (1974) coefficient for tests that can be divided into three congeneric part-tests with unknown lengths. Additionally, Gilmer and Feldt (1983) presented a more general formula applicable to tests with more than three congeneric parts with unknown lengths—it is referred to as the Feldt–Gilmer coefficient in Feldt and Brennan (1989). A maximum likelihood approach is also available, as described by Jöreskog (1971), through the use of a computer program LISREL (Jöreskog & Sörbom, 2018).

The internal consistency coefficients discussed thus far are suitable for tests consisting of items or part-tests that largely conform to the requirements of the essential tau-equivalence or congeneric models. Consider, however, a test composed of multiple content categories (or clusters), each with a distinct set of items. While items within each content category may satisfy the assumptions of the essential tau-equivalence or congeneric model, applying these assumptions to items across content categories may lead to violations. This is because some unique true score variance may be associated with each content category. The approach discussed in the next section employs a general framework for linear composites to address situations where parallel-form assumptions hold reasonably well within clusters but not across them.

Reliability of Linear Composites

A linear composite is defined as a weighted sum of scores from multiple constituent components: $Z = \sum_{m=1}^{M} w_m X_m$, where w_m is the weight associated with component m and M is the number of components. In this definition, weights are unconstrained—they can be positive, negative, or even zero. The general framework of linear composites considered in this section does not require that the constituent components in a composite are parallel in the classical sense of parallelism. In effect, a composite can be derived from any types of components. Consider the following examples: a gain score, which is the difference between pretest and posttest scores; battery composites, which are usually formed by averaging multiple subtest scores (e.g., English, math, reading, and science); predicted scores based on a multiple linear regression; and total test scores obtained by summing scores from several distinct item types.

Conceptually, for a linear composite, each instance of a replication would involve the same set of components with items within components that vary. Each individual component, X_m, may consist of a set of items that conform, more or less, to one of the CTT definitions of parallelism and is allowed to have its own true, error, and observed score variances and, consequently, reliability. If independent errors across components are assumed, the error variance for composite scores is simply a weighted sum of all component error variances: $\sigma^2(E_Z) = \sum w_m^2 \sigma^2(E_{X_m})$. This error variance for the composite is incorporated in the error-variance definition of reliability to yield a general-purpose reliability coefficient for a linear composite:

$$R(Z)_{EV} = 1 - \frac{\sigma^2(E_Z)}{\sigma^2(Z)} = 1 - \frac{\sum w_m^2 \sigma^2(E_{X_m})}{\sigma^2(Z)} = 1 - \frac{\sum w_m^2 \sigma^2(X_m)[1 - R(X_m)]}{\sigma^2(Z)}, \qquad (15)$$

where $R(X_m)$ is the reliability coefficient for component m. Equation 15 is not only convenient to use as long as a reliability estimate is available for each component, but also serves as a basis for deriving a reliability coefficient for many different types of linear composites. One such example is discussed next.

Stratified coefficient alpha (Rajaratnam et al., 1965) is generally considered more appropriate than the unstratified version, ${}_{\alpha}R(X)$, when a test is organized according to a table of content (or other type) specifications. The assumption of essential tau equivalence may be sensible within content categories (called "strata" more generally), but may be questionable across them because there is presumably some unique true score variance associated with each category. For estimation purposes, stratum-level scores are computed even though they may never be reported. Then, as a special case of Equation 15, stratified coefficient alpha is given by

$$_{strat\ \alpha}R(Z) = 1 - \frac{\sum \sigma^2(X_m)[1 - {}_{\alpha}R(X_m)]}{\sigma^2(Z)}. \tag{16}$$

The numerator in Equation 16 is the error variance in $_{strat\ \alpha}R(Z)$, which is the sum of error variances in the regular coefficient alpha associated with the strata. Occasionally, unstratified $_{\alpha}R(X)$ is applied to stratified data without explicitly taking into account stratification. Doing so will likely yield an underestimate of reliability, although the extent of the problem depends on the nature of strata. Stratification by content or item types often makes meaningful differences.

The formula presented in Equation 15 is "general" in the sense that it does not specify which reliability coefficient should be used for the constituent components. If $_{\alpha}R(X)$ is used for all components, the result is stratified alpha. In a similar vein, a stratified version of Feldt's coefficient may be obtained if one is willing to assume a congeneric model for the components. However, as W. Lee et al. (2025) demonstrated, stratification makes little difference between the stratified and unstratified Feldt's coefficients. As a final note, nothing in theory excludes the possibility of using different reliability coefficients for various components, although it is rarely done in practice.

RELIABILITY IN GT

It is undoubtedly true that the very simple model of CTT provides an accessible framework for addressing various measurement problems. However, a principal limitation of the simple model is the presence of only one error term (E), where all sources of measurement error are clumped together and cannot be disentangled. As an extension and liberalization of CTT (Brennan, 2001c; Cronbach et al., 1972; Feldt & Brennan, 1989), GT permits the explicit modeling of multiple sources of random error. This feature makes GT a powerful and flexible tool for studying reliability and related issues. The multifacet aspect of GT is achieved through the application of analysis of variance (ANOVA) procedures. Other important characteristics that distinguish GT from CTT include the following: (a) GT incorporates a conceptual framework that differentiates between the concepts of *universes of admissible observations* (and associated G studies) and *universes of generalization* (and associated D studies); (b) in GT, a single notion of *randomly parallel* forms replaces the multiple definitions of parallelism in CTT; (c) GT introduces a clear distinction between two different types of error—absolute error and relative error; and (d) multivariate GT provides an even broader framework for mixed models, including both random and fixed facets, with multiple universes of generalization.

The first comprehensive exploration of GT was presented in the book titled *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles* by Cronbach et al. (1972). However, the application of ANOVA to measurement problems predates the formalization of GT by Cronbach's team. Earlier works by researchers such as Burt (1936, 1955), Ebel (1951), Hoyt (1941), and Lindquist (1953, chap. 16) had already introduced the idea. A more accessible exposition of the theory is provided by Shavelson and Webb (1991). Brennan's (2001c) book, titled *Generalizability Theory,* has been widely cited, serving as both a textbook and a resource for students and practitioners. For a comprehensive review of the history of GT up to the present, readers are directed to Brennan (2022) and the references therein.

We shall now delve into some fundamental concepts, definitions, and assumptions, primarily focusing on univariate GT. Following this, there will be a brief overview of multivariate GT, and the section will conclude with an introduction to extended multivariate GT. While a full treatment of GT is beyond the scope of this section, we will use a few hypothetical scenarios to elucidate some crucial concepts and methods.

Univariate GT

Univariate GT provides a framework to model and estimate distinct sources of random error separately. To draw a comparison with CTT, a tautological model for univariate GT can be expressed as

$$X_{pf} = \pi_p + (E_{pf1} + E_{pf2} + \cdots E_{pfH}), \tag{17}$$

where π_p is the universe score for person p in the mean-score metric, which is analogous to the true score in CTT, and H is the number of sources of random error associated with "facets" (e.g., items, raters). The H error terms are closely tied to the investigator's intended universe of generalization (defined later), where replications $(f = 1, \ldots, \infty)$ are considered as comprising randomly parallel forms that differ with respect to the conditions of each facet. In this model, what constitutes error is a matter of definition and is emphatically under the control of the investigator.

Univariate GT is largely a two-step enterprise. The first step revolves around the notion of universes of admissible observations and generalizability studies (G studies), while the subsequent step considers the notion of universes of generalization and decision studies (D studies). These conceptual issues tend to be more challenging to grasp than the statistical issues concerning estimation of variance components and reliability statistics.

Universes of Admissible Observations and G Studies

Designing a measurement procedure begins with specifying a set of measurement conditions, known in GT as *facets*. To illustrate, consider a practical example of an English writing assessment. Among many possible considerations, one can identify facets of interest, such as writing prompts (denoted i) and raters (denoted r). Note that the investigator has full control over what and how many facets to consider given the specific purpose of the assessment. Both the numbers of potential writing prompts and the raters may be assumed to be indefinitely large, approaching infinity, at least theoretically. In this scenario, both facets are infinite in the *universe of admissible observations* (UAO), and the corresponding statistical model is called a random effects model, which is the primary focus of discussion in this section.[3] Furthermore, if any prompt in the universe could be evaluated by any rater in the universe, the two facets in the UAO are crossed, denoted $i \times r$. Now suppose there is a target *population* of persons (p) for whom the writing assessment is intended. In this context, persons are the *objects of measurement* about whom decisions are drawn. If any prompt and any rater in their universes can be associated with any person in the population, it is symbolized as $p \times i \times r$, which characterizes any observable data for the population and universes.

A linear model associated with the $p \times i \times r$ structure that represents any observed score for a single person on a single prompt evaluated by a single rater is expressed as

$$X_{pir} = \mu + \nu_p + \nu_i + \nu_r + \nu_{pi} + \nu_{pr} + \nu_{ir} + \nu_{pir}, \tag{18}$$

where μ is the grand mean in the population and universes and the ν terms are mutually uncorrelated *effects*. The term ν_{pir} in Equation 18 represents a three-way interaction combined with all other residual effects. It follows that the variance of observed scores for a *single* condition of each facet can be decomposed into seven uncorrelated *variance components*:

$$\sigma^2(X_{pir}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(r) + \sigma^2(pi) + \sigma^2(pr) + \sigma^2(ir) + \sigma^2(pir). \tag{19}$$

These variance components play a central role in all subsequent analyses. A study is conducted to collect data for the purpose of estimating these variance components for the UAO. Such a study is called a generalizability study, or more succinctly, a G study. One possible data collection design would involve a sample of n_p persons, each of whom responds to the same set of n_i prompts that are evaluated by the same set of n_r raters. This is called a $p \times i \times r$ G-study design, which is consistent with the structure of the UAO.

However, the UAO does not dictate a particular G-study design for collecting observed data. For example, if each of n_r raters provides ratings on the responses to two or more nonoverlapping prompts, this is a design symbolized as $p \times (i:r)$, where prompts are "nested" within raters. While there is no definitive answer to which G-study design should be used, it is generally preferred to use a crossed design (i.e., $p \times i \times r$) when the UAO is crossed. Using a crossed design provides the benefit that all seven variance components in Equation 19 can be independently estimated, offering great flexibility in obtaining results for other nested designs as well as the crossed design. By contrast, using a nested G-study design, some variance components in Equation 19 are not separately estimable (i.e., confounded), and results can only be obtained for certain nested designs but not for the crossed design.

Based on the G-study data, estimates of variance components can be obtained using various methods, including the ANOVA procedure. This procedure essentially involves the explicit use of the expected mean-square equations for a given G-study design, setting them equal to the weighted sum of contributing variance components. The expected mean squares are substituted with observed mean squares in the equations, and they are then solved for each of the variance components in terms of the mean squares. Once estimates of the variance components are obtained, they can be used to estimate error variances and reliability-like coefficients in subsequent decision studies, as discussed next.

Universes of Generalization and D Studies

As indicated in Equation 19, the G-study variance components for the UAO pertain to scores on *single* conditions of each facet. These estimated G-study variance components can be used in designing efficient measurement procedures for operational purposes through various decision (D) studies. Conceivably, an operational measurement procedure might encompass a set of measurement conditions (e.g., four prompts and two raters) from the UAO, and decisions about objects of measurement (persons) will be based on their *mean* (i.e., average) scores over those measurement conditions. Moreover, there exists a large (presumably infinite) number of similar forms that can be selected from the infinite UAO, each form containing different samples of measurement conditions (i.e., different sets of four prompts and two raters). Such forms are referred to as *randomly parallel forms,* which embody the notion of replications in GT. Unlike the definitions of parallelism in CTT, no constraints such as equal means and variances are imposed on randomly parallel forms, which are essentially indistinguishable from one another according to the principle of random sampling.

The concept of replications tied to randomly parallel forms is intricately linked to the notion of *universe of generalization* (UG). Indeed, a UG constitutes the universe of randomly parallel forms to which the investigator intends to generalize the results based on a particular form. In essence, a UAO is a universe for single observations, whereas a UG is a test-level universe focusing on mean scores over sets of measurement conditions. In a similar vein, the G-study variance components are derived from single-observation data and are subsequently employed to obtain the D-study variance components for mean scores over sets of conditions. The design structure of the D study may or may not mirror that of the G study. Furthermore, it is entirely legitimate for a UG to contain either all or a subset of conditions present in the UAO. These conceptual considerations are illustrated using the same example mentioned earlier.

Consider a scenario where the UAO is infinite with the $p \times i \times r$ structure, and G-study variance components have been estimated using the $p \times i \times r$ design with n_i prompts and n_r raters. Now, consider the following three different scenarios for D studies:

1. Same design and same universe as G study
   - Design: $p \times I \times R$
   - UG: both I and R are random (i.e., contains all the conditions in the UAO)
2. Same design but different universe
   - Design: $p \times I \times R$
   - UG: I is random, but R is a fixed facet (i.e., interest focuses on generalizing results over prompts only, but not raters)
3. Different design with same universe
   - Design: $p \times (I : R)$
   - UG: both I and R are random

Note that uppercase letters represent facets in D studies (i.e., I and R) to emphasize the use of *mean* scores over conditions, as opposed to the single prompt-rater scores in the G study. In addition, the D-study sample sizes for the prompts and raters can be user defined and may be the same as or different from the G-study sample sizes, denoted n'_i and n'_r.

It is particularly important to note that the specification of a UG and characterization of a D-study design bear resemblance to the *conceptualization-estimation* framework for reliability discussed in the section "Terms and Background." To define a UG, an investigator needs to articulate or conceptualize what constitutes a replication of a measurement procedure, predominantly in terms of which facets are considered random and which are fixed. The features of a D-study design pertain to estimation of reliability statistics.

The linear model for the first D-study scenario (i.e., $p \times I \times R$ with both I and R random) representing a person's observed mean score over n'_i prompts and n'_r raters can be expressed as

$$X_{pIR} = \overline{X}_p = \mu + \nu_p + \nu_I + \nu_R + \nu_{pI} + \nu_{pR} + \nu_{IR} + \nu_{pIR}, \tag{20}$$

which is analogous to Equation 18 except for the use of uppercase subscripts, I and R. The D-study variance components for the score effects in Equation 20 can be obtained by dividing the corresponding G-study variance components by the user-defined D-study sample sizes. In this example, $\sigma^2(I) = \sigma^2(i)/n'_i$, $\sigma^2(R) = \sigma^2(r)/n'_r$,

$\sigma^2(pI) = \sigma^2(pi)/n'_i$, $\sigma^2(pR) = \sigma^2(pr)/n'_r$, $\sigma^2(IR) = \sigma^2(ir)/n'_i n'_r$, and $\sigma^2(pIR) = \sigma^2(pir)/n'_i n'_r$.

As noted earlier, GT employs the expected-value perspective on true score, which it refers to as *universe score*. That is, universe score is defined as the expected value of observed mean scores over all randomly parallel forms in the UG:

$$\pi_p \equiv \mu_p \equiv \mathscr{E}_T \mathscr{E}_R X_{pIR}. \tag{21}$$

The variance of universe scores over all persons in the population is defined as $\sigma^2(\pi) = \mathscr{E}_p\left(\mu_p - \mu\right)^2$. For the $p \times I \times R$ design with random I and R, the universe score variance is simply $\sigma^2(\pi) = \sigma^2(p)$.

The second scenario differs from the first in that its UG is more narrowly defined than the UAO by fixing the rater facet. In this case, every randomly parallel form consists of the exact same set of raters, eliminating generalization over raters. The net effect is that the variance component for the person-by-rater interaction effect contributes to universe score variance; specifically, $\sigma^2(\pi) = \sigma^2(p) + \sigma^2(pR)$.

For the third scenario with a $p \times (I : R)$ D-study design, the linear model is

$$X_{pIR} = \overline{X}_p = \mu + \nu_p + \nu_R + \nu_{I:R} + \nu_{pR} + \nu_{pI:R}. \tag{22}$$

This linear model has fewer score effects (and variance components) compared to the $p \times I \times R$ design shown in Equation 20. If the G-study variance components for the fully crossed $p \times i \times r$ design are available, another G study is not required to obtain the D-study variance components for the nested design. Instead, a nested effect can be expressed in terms of confounded effects from the crossed design; that is, $\sigma^2(i : r) = \sigma^2(i) + \sigma^2(ir)$ and $\sigma^2(pi : r) = \sigma^2(pi) + \sigma^2(pir)$. It follows that $\sigma^2(I:R) = \left[\sigma^2(i) + \sigma^2(ir)\right]/n'_i n'_r$ and $\sigma^2(pI:R) = \left[\sigma^2(pi) + \sigma^2(pir)\right]/n'_i n'_r$ — the other three variance components are the same for both designs. Although the design for the third scenario differs from the first one, universe score variance remains unchanged (i.e., $\sigma^2(p)$) because the UG is the same.

Error Variances and Coefficients

Reliability coefficients in CTT are fundamentally correlation-based, directly or indirectly focusing on the rank ordering of test takers. Consequently, error variance associated with CTT reliability coefficients is interpreted in relation to the performance of other persons in the group, making it inherently *relative*. By contrast, GT distinguishes between two types of error: absolute error and relative error, each with different interpretations and uses. Absolute error for a person is the difference between their observed mean score and the universe score:

$$\Delta_p \equiv \overline{X}_p - \mu_p. \tag{23}$$

The variance of Δ_p over persons is referred to as *absolute error variance,* denoted $\sigma^2(\Delta)$, which is more suitable for criterion-referenced interpretations of scores. On the contrary, relative error, akin to CTT, is the difference between observed deviation scores and universe deviation scores:

$$\delta_p \equiv \left(\bar{X}_p - \mathscr{E}_p\bar{X}_p\right) - \left(\mu_p - \mu\right). \tag{24}$$

Relative error variance, $\sigma^2(\delta)$, is the variance of Equation 24 over persons.

Each type of error variance has an associated reliability-like coefficient. A generalizability coefficient involves relative error variance as its error term, defined as

$$\mathscr{E}\rho^2 = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\delta)}. \tag{25}$$

A different coefficient that involves absolute error variance is called a dependability coefficient, given by

$$\Phi = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\Delta)}. \tag{26}$$

Results for error variances and coefficients depend on a specified UG, a D-study design, and corresponding D-study variance components. Relative error variance, $\sigma^2(\delta)$, and a generalizability coefficient, $\mathscr{E}\rho^2$, are analogous to error variance and a reliability coefficient, respectively, in CTT. For a simple $p \times I$ design, $\mathscr{E}\rho^2$ is equal to coefficient alpha (or KR20 for binary items) if $n'_i = n_i$; however, this equality does not extend to other UGs and/or D-study designs.

It can be shown that absolute error variance is the sum of all variance components except $\sigma^2(p)$ for a random model with a crossed design. In the first D-study scenario with the $p \times I \times R$ design (both I and R random),

$$\sigma^2(\Delta) = \sigma^2(I) + \sigma^2(R) + \sigma^2(pI) + \sigma^2(pR) + \sigma^2(IR) + \sigma^2(pIR).$$

Relative error variance, by contrast, contains all variance components for interaction effects that involve p:

$$\sigma^2(\delta) = \sigma^2(pI) + \sigma^2(pR) + \sigma^2(pIR).$$

Clearly, $\sigma^2(\delta)$ is smaller than $\sigma^2(\Delta)$, which, in turn, leads to $\mathscr{E}\rho^2$ that is larger than Φ.

In the second scenario with R fixed, $\sigma^2(pR)$ contributes to universe score variance, resulting in

$$\sigma^2(\Delta) = \sigma^2(I) + \sigma^2(pI) + \sigma^2(IR) + \sigma^2(pIR)$$

$$\sigma^2(\delta) = \sigma^2(pI) + \sigma^2(pIR).$$

Compared to the first scenario, this leads to larger universe score variance, smaller error variances, and larger coefficients. This outcome is a consequence of the reduced error in a restricted UG that limits the generalization of results.

For the third scenario, involving an infinite UG with a $p \times (I : R)$ design,

$$\sigma^2(\Delta) = \sigma^2(R) + \sigma^2(I : R) + \sigma^2(pR) + \sigma^2(pI : R)$$
$$\sigma^2(\delta) = \sigma^2(pR) + \sigma^2(pI : R)\cdot$$

Results for this design are likely to differ from those of the first scenario with a fully crossed design, even if both share the same UG. This underscores the dependence of D-study results on both UG characteristics and D-study design structure.

The hypothetical scenarios with two-facet designs served as a vehicle to introduce key concepts, terminology, and statistical formulas in univariate GT. They also provided insight into the considerations necessary for establishing a well-defined measurement procedure. Perhaps the most critical consideration, after identifying all pertinent sources of error (i.e., facets), is determining which facets are random and which are fixed in the intended UG. Other factors demanding attention include the D-study design structure and D-study sample sizes. All these considerations affect D-study results, one way or another.

Multivariate GT

An early exploration of multivariate GT was presented by Rajaratnam et al. (1965), particularly in the context of introducing stratified alpha as a tool for estimating reliability for a composite. Considered as a precursor to multivariate GT, stratified alpha set the stage for later developments. Later, Cronbach et al. (1972) offered the first integrated treatment of multivariate GT, a framework further expanded by Brennan (2001c).

Expanding on the univariate model depicted in Equation 17, multivariate GT may be represented as

$$X_{pf} = \left(\pi_{p1} + \pi_{p2} + \cdots + \pi_{pM}\right) + \left(E_{pf1} + E_{pf2} + \cdots + E_{pfH}\right), \qquad (27)$$

in which each person possesses M universe scores, each associated with one of M fixed conditions of measurement; and, as previously defined, the number of sources of error (i.e., facets), denoted by H, is typically the same across all M levels of a fixed facet—more intricate multivariate designs are considered in the section "Extended Multivariate GT." While many fundamental principles of univariate GT still apply to multivariate GT, a key distinction lies in the presence of "multiple" univariate random effects models, each corresponding to each level of the fixed facet.

To apply multivariate GT, there must be at least one fixed facet and at least one random facet in the UAO and G-study data collection design. For illustration, let us consider a simple example based on the so-called table of specifications model

(Jarjoura & Brennan, 1982, 1983). In this model, each item in the UAO is associated with one of several specific content categories, making items (i) random and content categories (c) fixed. Suppose a G study is conducted to gather response data from a group of test takers (p) who respond to all items nested within each content category. This multivariate design is designated $p^{\bullet} \times i^{\circ}$, where the filled circle superscript indicates that the facet is crossed with the levels of the fixed facet, c, and the empty circle superscript signifies the facet nested within c.

For simplicity, suppose there are two content categories (i.e., $n_c = 2$). Under the multivariate $p^{\bullet} \times i^{\circ}$ design, covariances exist for the object of measurement (persons) since the same persons respond to all items in both content categories. However, there are no covariances for items that are nested within c. The variance and covariance components for p, i, and pi are displayed in the following $n_c \times n_c$ matrices:

$$\boldsymbol{\Sigma}_p = \begin{bmatrix} \sigma_1^2(p) & \sigma_{12}(p) \\ \sigma_{12}(p) & \sigma_2^2(p) \end{bmatrix}, \boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_1^2(i) & \\ & \sigma_2^2(i) \end{bmatrix}, \text{ and } \boldsymbol{\Sigma}_{pi} = \begin{bmatrix} \sigma_1^2(pi) & \\ & \sigma_2^2(pi) \end{bmatrix},$$

where the subscripts 1 and 2 denote the two content categories. Note that both $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}_{pi}$ are diagonal matrices with zero covariances. Since there is a univariate random $p \times i$ design for each content level, the variances reported in the first and second columns in the matrices represent the variance components for a $p \times i$ design based solely on data from the first and second content categories, respectively. Hence, the conventional ANOVA procedure can be used to estimate univariate variance components for each content category. The estimate of the covariance component in $\boldsymbol{\Sigma}_p$, $\sigma_{12}(p)$, is simply the covariance between persons' observed scores on the two categories. For more complex designs, interested readers can refer to Brennan (2001c, chaps. 9 and 11) for procedures on estimating covariance components.

Now, consider a subsequent D study that involves n'_{i1} items for Category 1 and n'_{i2} items for Category 2 with the same design structure as the G study. In this case, the two diagonal elements in the variance–covariance component matrix for I, $\boldsymbol{\Sigma}_I$, are $\sigma_1^2(I) = \sigma_1^2(i)/n'_{i1}$ and $\sigma_2^2(I) = \sigma_2^2(i)/n'_{i2}$. Likewise, $\boldsymbol{\Sigma}_{pI}$ contains the following diagonal elements: $\sigma_1^2(pI) = \sigma_1^2(pi)/n'_{i1}$ and $\sigma_2^2(pI) = \sigma_2^2(pi)/n'_{i2}$. It follows that the matrices for universe score, relative error, and absolute error, respectively, are $\boldsymbol{\Sigma}_\pi = \boldsymbol{\Sigma}_p$; $\boldsymbol{\Sigma}_\delta = \boldsymbol{\Sigma}_{pI}$; and $\boldsymbol{\Sigma}_\Delta$ with diagonal elements of $\sigma_1^2(\Delta) = \sigma_1^2(I) + \sigma_1^2(pI)$ and $\sigma_2^2(\Delta) = \sigma_2^2(I) + \sigma_2^2(pI)$. Note that for this design, $\boldsymbol{\Sigma}_\Delta$ and $\boldsymbol{\Sigma}_\delta$ are diagonal matrices.

In practical applications, decisions about test takers are often based on their composite scores over all levels of content categories. A composite based on two content categories can be expressed as a weighted sum of two mean scores: $\overline{Z} = w_1\overline{X}_1 + w_2\overline{X}_2$, where the bar above Z denotes the mean-score metric, and weights are typically proportional to the number of items in each category (i.e., $w_1 = n'_{i1}/(n'_{i1}+n'_{i2})$ and $w_2 = n'_{i2}/(n'_{i1}+n'_{i2})$). Similarly, the composite universe score for a person is defined as $\pi_{p\bar{Z}} = w_1\pi_{p1} + w_2\pi_{p2}$. Since $\boldsymbol{\Sigma}_\pi = \boldsymbol{\Sigma}_p$, composite universe score variance $\sigma^2_{\bar{Z}}(\pi)$ is the sum of all of the elements in $\boldsymbol{\Sigma}_p$ with associated weights: $\sigma^2_{\bar{Z}}(\pi) = w_1^2\sigma_1^2(p) + w_2^2\sigma_2^2(p) + 2w_1w_2\sigma_{12}(p)$. Relative error variance and absolute error variance for the composite scores, respectively, are weighted sums of diagonal elements of their corresponding matrices, as follows: $\sigma^2_{\bar{Z}}(\delta) = w_1^2\sigma_1^2(\delta) + w_2^2\sigma_2^2(\delta)$ and $\sigma^2_{\bar{Z}}(\Delta) = w_1^2\sigma_1^2(\Delta) + w_2^2\sigma_2^2(\Delta)$. These error variances, when combined with composite universe score variance, provide reliability-like coefficients for the composite scores. A multivariate generalizability coefficient and a multivariate dependability coefficient for the composite are defined similarly to Equations 25 and 26, respectively.

The typical computational procedures for multivariate GT designs have been demonstrated using a simple $p^\bullet \times i^\circ$ design so far. Applying these procedures to multifacet situations and/or different design structures is straightforward as long as the multivariate design is properly identified. For example, consider a scenario with two types (c) of short free-response items (i) that are evaluated by raters (r), where each rater scores test takers' (p) responses to a *different* subset of items for both types. This constitutes a multivariate $p^\bullet \times (i^\circ : r^\bullet)$ design with c serving as a multivariate fixed facet. Items are nested within both r and c, and for each level of c, there is a univariate random $p \times (i : r)$ design. Some effects (p, r, and pr) will have full matrices with nonzero covariance terms, while other effects ($i : r$ and $pi : r$) will have diagonal matrices with zero covariances. D-study variance and covariance components, along with various D-study statistics, can then be computed using the procedures delineated in the preceding discussion of the simplest design.

The procedures discussed thus far are generally sufficient for most applications with a "single" multivariate design where the same univariate design structure applies to all levels of the fixed facet. Nonetheless, more complex designs can arise in practical situations, necessitating special treatment, as discussed in the following section.

Extended Multivariate GT

The extant literature on GT offers limited guidance on handling a mixture of different design structures in a multivariate design. One such example is a mixed-format exam that contains both multiple-choice (MC) items and free-response (FR) items involving raters for scoring. Here, the item type represents a multivariate fixed facet with two levels, MC and FR. Moses and Kim (2015) were the first to consider combining two distinct designs in multivariate GT, and Brennan et al. (2022) later provided a more extensive treatment, referring to it as *extended* multivariate GT.

The model representation for conventional multivariate GT shown in Equation 27 may be further expanded to characterize extended multivariate GT. Specifically,

$$X_{pf} = \left[\pi_{p1} + (E_{pf1} + \cdots + E_{pfH_1})\right] + \cdots + \left[\pi_{pM} + (E_{pf1} + \cdots + E_{pfH_M})\right]. \quad (28)$$

This model suggests that there are M distinct levels of a fixed facet, each associated with its own sources of error (i.e., facets) signified by different subscripts to H, H_1 through H_M. Since this model has not been fully exploited in the literature, one of the examples reported by Brennan et al. (2022) is briefly introduced here.

To continue with the example of a mixed-format exam, the MC section involves items (i) as the primary measurement condition, while the FR section involves both items (i) and raters (r). For notational simplicity, the symbol i is used to represent both MC and FR items—however, it does not imply interchangeability between the two item types. Representing the combination of two universes, the UAO structure can be symbolized as $\{p^{\bullet} \times i^{\circ}\}\{p^{\bullet} \times i^{\circ} \times r^{\circ}\}$, where persons are crossed with both universes, as indicated by the closed circle superscript. Brennan et al. (2022) concisely represented this design as $p^{\bullet} \times \left[i^{\circ} \cup \left(i^{\circ} \times r^{\circ}\right)\right]$, where the symbol $\cup$ denotes the union of the MC and FR sections, and the first and second i° before and after $\cup$ are associated with the MC and FR sections, respectively.

Suppose G-study data were collected for the $p^{\bullet} \times \left[i^{\circ} \cup \left(i^{\circ} \times r^{\circ}\right)\right]$ design. The conventional univariate methods can be applied to estimate G-study variance components separately for the MC and FR sections based on a univariate $p \times i$ design for the MC section and a univariate $p \times i \times r$ design for the FR section. For the full multivariate design, the covariance component for p needs to be estimated. Given that p is the only linked facet in this design, an unbiased estimate of the covariance is simply the observed covariance between MC and FR scores.

The left side of Table 5.1 presents the G-study variance and covariance components for this multivariate design. Each matrix reports variance components for MC and FR (indicated by the subscripts 1 and 2) in the first and second columns, respectively. Notably, the MC section, analyzed with a $p \times i$ design, involves three relevant variance components $(p, i$, and $pi)$. In contrast, the FR section, analyzed with a $p \times i \times r$ design, includes seven variance components (p, i, r, pi, pr, ir, and pir). Each of the seven matrices contains variance components corresponding to one of the seven effects for FR. However, only three matrices ($\boldsymbol{\Sigma}_p$, $\boldsymbol{\Sigma}_i$, and $\boldsymbol{\Sigma}_{pi}$) report variance components for MC. This is because the other four matrices are not relevant to MC, as indicated by "NA." Covariance components are present in $\boldsymbol{\Sigma}_p$ only.

On the right side of Table 5.1, the D-study variance and covariance components are presented. These components are obtained by dividing their corresponding G-study components by the user-defined D-study sample sizes, n'_{i1}, n'_{i2}, and n'_r. Composite universe score variance is calculated as the weighted sum of all the elements in $\boldsymbol{\Sigma}_p$: $\sigma^2_{\bar{Z}}(\pi) = w_1^2\sigma_1^2(p) + w_2^2\sigma_2^2(p) + 2w_1w_2\sigma_{12}(p)$. Since MC and FR involve different sources of error, both relative error variance and absolute error variance for the

Table 5.1 Variance and Covariance Components for the Extended Multivariate Design Example

G study	D study
$p^{\bullet} \times \left[i^{\circ} \cup \left(i^{\circ} \times r^{\circ} \right) \right]$	$p^{\bullet} \times \left[I^{\circ} \cup \left(I^{\circ} \times R^{\circ} \right) \right]$
MC FR	MC FR
$\boldsymbol{\Sigma}_p = \begin{bmatrix} \sigma_1^2(p) & \sigma_{12}(p) \\ \sigma_{12}(p) & \sigma_2^2(p) \end{bmatrix}$	$\boldsymbol{\Sigma}_p = \begin{bmatrix} \sigma_1^2(p) & \sigma_{12}(p) \\ \sigma_{12}(p) & \sigma_2^2(p) \end{bmatrix}$
$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_1^2(i) & \\ & \sigma_2^2(i) \end{bmatrix}$	$\boldsymbol{\Sigma}_I = \begin{bmatrix} \dfrac{\sigma_1^2(i)}{n'_{i1}} & \\ & \dfrac{\sigma_2^2(i)}{n'_{i2}} \end{bmatrix}$
$\boldsymbol{\Sigma}_r = \begin{bmatrix} NA & \\ & \sigma_2^2(r) \end{bmatrix}$	$\boldsymbol{\Sigma}_R = \begin{bmatrix} NA & \\ & \dfrac{\sigma_2^2(r)}{n'_r} \end{bmatrix}$
$\boldsymbol{\Sigma}_{pi} = \begin{bmatrix} \sigma_1^2(pi) & \\ & \sigma_2^2(pi) \end{bmatrix}$	$\boldsymbol{\Sigma}_{pI} = \begin{bmatrix} \dfrac{\sigma_1^2(pi)}{n'_{i1}} & \\ & \dfrac{\sigma_2^2(pi)}{n'_{i2}} \end{bmatrix}$
$\boldsymbol{\Sigma}_{pr} = \begin{bmatrix} NA & \\ & \sigma_2^2(pr) \end{bmatrix}$	$\boldsymbol{\Sigma}_{pR} = \begin{bmatrix} NA & \\ & \dfrac{\sigma_2^2(pr)}{n'_r} \end{bmatrix}$
$\boldsymbol{\Sigma}_{ir} = \begin{bmatrix} NA & \\ & \sigma_2^2(ir) \end{bmatrix}$	$\boldsymbol{\Sigma}_{IR} = \begin{bmatrix} NA & \\ & \dfrac{\sigma_2^2(ir)}{n'_{i2}\,n'_r} \end{bmatrix}$

$\Sigma_{pir} = \begin{bmatrix} NA & \\ & \sigma_2^2(pir) \end{bmatrix}$	$\Sigma_{pIR} = \begin{bmatrix} NA & \\ & \dfrac{\sigma_2^2(pir)}{n'_{i2}n'_r} \end{bmatrix}$

Note. D study = decision study; FR = free response; G study = generalizability study; MC = multiple choice.

composite involve different variance and covariance components for the two sections, as follows:

$$\sigma_{\bar{Z}}^2(\delta) = w_1^2\left[\sigma_1^2(pI)\right] + w_2^2\left[\sigma_2^2(pI) + \sigma_2^2(pR) + \sigma_2^2(pIR)\right], \tag{29}$$

and

$$\begin{aligned}\sigma_{\bar{Z}}^2(\Delta) = w_1^2\left[\sigma_1^2(I) + \sigma_1^2(pI)\right] &+ w_2^2[\sigma_2^2(I) + \sigma_2^2(R) + \sigma_2^2(IR) \\ &+ \sigma_2^2(pI) + \sigma_2^2(pR) + \sigma_2^2(pIR)].\end{aligned} \tag{30}$$

Finally, the reliability-like coefficients are obtained using these quantities for the composite in Equations 25 and 26.

The preceding discussion highlights that extended multivariate GT offers a useful framework for dealing with multivariate designs that are more complex than those commonly discussed in the current literature. However, the model needs to be studied further through more diversified applications of the model with various types of complex designs in real-world testing.

RELIABILITY IN IRT

In IRT, a test taker's response to an item is modeled as a function of an underlying proficiency variable denoted θ and item parameters. The proficiency for a test taker can be estimated using either the maximum likelihood (ML) or Bayesian estimation methods. An estimator, denoted $\hat{\theta}$, serves as the observed score in IRT. In this context, we primarily consider the ML and Bayesian expected a posteriori (EAP) estimators. Another Bayesian estimator, maximum a posteriori (MAP), is briefly discussed because it shares the same underlying framework with the EAP estimator.

The main focus of this section is to present various approaches to quantifying reliability within the framework of IRT. The mathematical formulations presented here are designed to be applicable and generalizable to many different IRT models, without making specific distinctions among those models. Throughout this section, unidimensional dichotomous IRT models are assumed, except in the last subsection where some procedures are discussed for a few specific multidimensional IRT (MIRT) models.

Results are presented for several score metrics, including ML estimates, EAP estimates, summed raw scores, transformed scale scores from summed raw scores, composite scores, and composite scale scores. The two classes of estimation methods (ML and EAP) are discussed separately because of their fundamental inconsistencies in terms of the model framework and assumptions.

Under the ML framework, the data at hand are viewed as a random sample from a distribution with known θ and item parameters. For a test taker with θ, the observed data over replications are conceptually obtained by administering test forms with the exact same item parameters repeatedly. These forms, with the same item parameters, are often called *strictly parallel* forms, or it can be said that the form is effectively fixed (W. Lee et al., 2000). Therefore, the source of error causing the variability in estimates over replications is best described as resulting from each item being sampled from a large pool of items with identical parameters.

Reliability statistics under the ML framework are conceptualized and computed by focusing on the distribution of proficiency estimates conditional on the parameter $f(\hat{\theta}|\theta)$. The standard deviation of $\hat{\theta}$ given θ is termed the conditional standard error of estimation (SEE). In IRT, the SEE plays a role similar to that of the SEM in CTT. However, these terms are often used interchangeably in the context of estimating reliability in IRT (e.g., B. F. Green et al., 1984; Lord, 1983). For the purposes of this chapter, no distinction is made between the two terms.

In contrast, the Bayesian framework is primarily concerned with obtaining the posterior distribution of parameters given observed data, $f(\theta|\boldsymbol{u})$, where parameters are considered random while data are fixed. Here, $\boldsymbol{u}$ is a vector of a test taker's item responses. The EAP estimate is the mean of the posterior distribution of θ for test takers with the same item responses. The standard deviation of the posterior distribution serves as the SEE for the EAP estimator. However, as demonstrated later, under this Bayesian framework, the concept of replications is obscure, at best, and thus the Bayes SEE is conceptually different from the SEE in the ML estimator and SEM in CTT.

Consider the following parameterization, similar to CTT:

$$\hat{\theta} = \theta + \varepsilon, \tag{31}$$

where a proficiency estimate $\hat{\theta}$ is decomposed into the latent proficiency θ and an error component, ε. For simplicity, subscripts indicating test takers and forms are omitted here. In contrast to CTT, where the true score for a test taker is equal to the expected value of observed scores, the proficiency estimators, both ML and EAP, provide biased estimates for θ (J. K. Kim & Nicewander, 1993; Lord, 1983; Warm, 1989). The ML estimator is unbiased only asymptotically as the test length approaches infinity. The bias in the ML estimates has significant consequences. Unlike the results in CTT, various expressions or definitions of the reliability coefficient for the ML estimator may not yield the same answer, a fact that is frequently overlooked.

Maximum Likelihood Proficiency Estimates

Let $\widehat{\theta}$ denote the ML estimator of θ. As noted earlier, due to bias in $\widehat{\theta}$, the expected value of $\widehat{\theta}$ is not equal to θ; namely, $\mathscr{E}(\widehat{\theta}|\theta) \neq \theta$. The expected value of $\widehat{\theta}$ is referred to here as *expected* proficiency, compared to *latent* proficiency θ. For $\widehat{\theta}$, the following reliability coefficients are typically considered:

$$R(\widehat{\theta})_{PF} \equiv \rho\left(\widehat{\theta}, \widehat{\theta}'\right); R(\widehat{\theta})_{SC} \equiv \rho^2(\widehat{\theta}, \theta); R(\widehat{\theta})_{VR} \equiv \frac{\sigma^2(\theta)}{\sigma^2(\widehat{\theta})};$$

$$R(\widehat{\theta})_{EVR} \equiv \frac{\sigma^2[\mathscr{E}(\widehat{\theta}\,|\,\theta)]}{\sigma^2(\widehat{\theta})} = 1 - \frac{\mathscr{E}\sigma^2(\widehat{\theta}\,|\,\theta)}{\sigma^2(\widehat{\theta})}; \text{ and } R(\widehat{\theta})_{MVR} \equiv \frac{\sigma^2(\theta)}{\sigma^2(\theta) + \mathscr{E}\sigma^2(\widehat{\theta}\,|\,\theta)}.$$

In CTT, similar expressions of these reliability coefficients are all identical when derived under the assumptions of classically parallel forms. However, the equality does not hold for $\widehat{\theta}$. The first definition $R(\widehat{\theta})_{PF}$ is the parallel-forms coefficient, $R(\widehat{\theta})_{SC}$ is the squared-correlation coefficient, and the last three represent the variance-ratio coefficients (indicated by subscripts including *VR*). These coefficients will likely produce different results primarily because they use different definitions of true proficiency and its variance. Hereinafter, $\sigma^2(\theta)$ in $R(\widehat{\theta})_{VR}$ and $R(\widehat{\theta})_{MVR}$ is referred to as *latent proficiency variance*, while the numerator of the first formula in $R(\widehat{\theta})_{EVR}$, $\sigma^2[\mathscr{E}(\widehat{\theta}|\theta)]$, is referred to as *expected proficiency variance*. The subscript *EVR* appended to $R(\widehat{\theta})_{EVR}$ signifies the use of expected proficiency variance. Likewise, $\mathscr{E}\sigma^2(\widehat{\theta}\,|\,\theta)$ in $R(\widehat{\theta})_{EVR}$ and $R(\widehat{\theta})_{MVR}$ is referred to here as *expected error variance*, which differs from either the error variance $\sigma^2(\varepsilon)$ (see Equation 31) or another quantity, $\sigma^2(\widehat{\theta}) - \sigma^2(\theta)$. Recall that $\mathscr{E}\sigma^2(\widehat{\theta}\mid\theta)$ quantifies the variability of $\widehat{\theta} - \mathscr{E}\widehat{\theta}$; by contrast, $\sigma^2(\varepsilon)$ is concerned about $\widehat{\theta} - \theta$. The last three variance-ratio coefficients differ in terms of which types of variances are used in the numerators and denominators. In effect, $R(\widehat{\theta})_{VR}$ involves latent proficiency variance in conjunction with $\sigma^2(\widehat{\theta}) - \sigma^2(\theta)$ as its error variance; $R(\widehat{\theta})_{EVR}$ contains expected variances in both the numerator and denominator; and $R(\widehat{\theta})_{MVR}$ exploits an interesting mix of both types (note the letter *M* in the subscript).

For ease of explanation, let's first delve into the three variance-ratio coefficients, followed by $R(\widehat{\theta})_{PF}$ and $R(\widehat{\theta})_{SC}$. In the literature, $R(\widehat{\theta})_{EVR}$ is often called marginal reliability (B. F. Green et al., 1984). S. Kim (2012) labeled $R(\widehat{\theta})_{EVR}$ as the (squared) correlation ratio for predicting θ from $\widehat{\theta}$, which might also be called an intraclass correlation coefficient based on the ANOVA identity, given by

$$\sigma^2(\widehat{\theta}) = \sigma^2[\mathscr{E}(\widehat{\theta}|\theta)] + \mathscr{E}\sigma^2(\widehat{\theta}|\theta). \tag{32}$$

The intraclass correlation coefficient, $R(\hat{\theta})_{EVR}$, becomes identical to the squared Pearson correlation coefficient (i.e., $R(\hat{\theta})_{SC}$) if the relationship between $\hat{\theta}$ and θ is linear; however, if the relationship is not linear, $R(\hat{\theta})_{EVR} \geq R(\hat{\theta})_{SC}$.

Since $\hat{\theta}$ is a biased estimator, $\mathscr{E}(\hat{\theta}|\theta) = \theta + BIAS$. Bias is not constant and can be either positive or negative depending on the location of a proficiency parameter along the θ scale. An immediate consequence of the bias in $\hat{\theta}$ is that $\sigma^2[\mathscr{E}(\hat{\theta}|\theta)] \neq \sigma^2(\theta)$. Rather,

$$\sigma^2[\mathscr{E}(\hat{\theta}|\theta)] = \sigma^2(\theta + BIAS) = \sigma^2(\theta) + \sigma^2(BIAS) + 2\sigma(\theta, BIAS), \quad (33)$$

where the variance terms are always positive and the covariance between θ and $BIAS$ is usually positive although there is no theoretical justification or proof for that (Lord, 1983). Clearly, the expected proficiency variance, $\sigma^2[\mathscr{E}(\hat{\theta}|\theta)]$, is *not* equal to $\sigma^2(\theta)$ and the former usually is larger than the latter. From Equations 32 and 33, it follows that

$$\sigma^2(\hat{\theta}) = \sigma^2(\theta) + \sigma^2(BIAS) + 2\sigma(\theta,BIAS) + \mathscr{E}\sigma^2(\hat{\theta}|\theta). \quad (34)$$

As $\sigma(\theta, BIAS)$ typically is positive, $\sigma^2(\hat{\theta}) > \sigma^2(\theta)$.

Coefficient $R(\hat{\theta})_{EVR}$ uses the expected proficiency variance in the numerator and the expected error variance in the denominator because $\sigma^2(\hat{\theta}) - \sigma^2[\mathscr{E}(\hat{\theta}|\theta)] = \mathscr{E}\sigma^2(\hat{\theta}|\theta)$. Conceptually, $BIAS = \mathscr{E}(\hat{\theta}|\theta) - \theta$ is a constant for a given test taker with θ (i.e., systematic error), and thus its variance (and covariance) should contribute to the true proficiency variance, but not to the error variance. Equations 33 and 34 clearly show that the expected proficiency variance, $\sigma^2[\mathscr{E}(\hat{\theta}|\theta)]$, in $R(\hat{\theta})_{EVR}$ incorporates two additional terms related to bias, in addition to $\sigma^2(\theta)$, whereas the expected error variance does not encompass them.

To estimate $R(\hat{\theta})_{E\,VR} = 1 - \left[\mathscr{E}\sigma^2(\hat{\theta}|\theta)/\sigma^2(\hat{\theta})\right]$, the conditional error variance, $\sigma^2(\hat{\theta}\,|\,\theta)$, is often obtained by taking the reciprocal of the test information function as $\sigma^2(\hat{\theta}|\theta) = 1\,/\,I(\theta,\hat{\theta})$. Thus, $R(\hat{\theta})_{EVR}$ can also be expressed as

$$R(\hat{\theta})_{EVR} = 1 - \frac{\mathscr{E}\sigma^2(\hat{\theta}|\theta)}{\sigma^2(\hat{\theta})} = 1 - \frac{\mathscr{E}[1/I(\theta,\hat{\theta})]}{\sigma^2(\hat{\theta})}. \quad (35)$$

Taking the three-parameter logistic IRT model as an example, the test information on an n-item test is simply the sum of the item information functions: $I(\theta,\hat{\theta}) = \sum_{i=1}^{n} I_i(\theta,\hat{\theta}) = \sum_{i=1}^{n} \left[\left[P_i'(\theta)\right]^2 / P_i(\theta)Q_i(\theta)\right]$, where $I_i(\theta,\hat{\theta})(i = 1,\ldots,n)$ is the item information functions, $P_i(\theta) = c_i + (1 - c_i)/\{1 + \exp[-1.7a_i(\theta - b_i)]\}$, and $Q_i(\theta) = 1 - P_i(\theta)$. The expectation in Equation 35 typically is taken over a set of discrete θ quadrature points and associated weights assuming a standard normal distribution.

Coefficient $R(\hat{\theta})_{VR}$ differs from $R(\hat{\theta})_{EVR}$ in its use of latent proficiency variance and a distinct form of error variance expressed as $\sigma^2(\hat{\theta}) - \sigma^2(\theta)$:

$$\sigma^2(\hat{\theta}) - \sigma^2(\theta) = \mathscr{E}\sigma^2(\hat{\theta}|\theta) + \sigma^2(BIAS) + 2\sigma(\theta,BIAS). \quad (36)$$

This "error variance" comprises the expected error variance plus two terms associated with systematic error, namely, bias. The bias-related terms, which contribute to the expected proficiency variance in $R(\hat{\theta})_{EVR}$, are integrated into the error variance in $R(\hat{\theta})_{VR}$, resulting in $R(\hat{\theta})_{EVR} \geq R(\hat{\theta})_{VR}$. This inclusion of bias terms in the error variance of $R(\hat{\theta})_{VR}$ raises concerns about the defensibility of this coefficient because systematic errors, in principle, should contribute to the true proficiency variance.

The final variance-ratio reliability coefficient, $R(\hat{\theta})_{MVR}$, is widely discussed and used in practice (e.g., Andersson & Xin, 2018; Cheng et al., 2012). Assuming the distribution of θ follows a standard normal distribution with a variance equal to 1, $R(\hat{\theta})_{MVR}$ can be expressed as $1/\{1 + \mathscr{E}[1/I(\theta, \hat{\theta})]\}$. The prevalent use of $R(\hat{\theta})_{MVR}$ relies on convenience—it does not require the use of sample estimates $\hat{\theta}$s, assumes the variance of θ to be 1, and easily obtains the information function using standard normal quadrature points and weights. Coefficient $R(\hat{\theta})_{MVR}$ is an interesting statistic because it uses a mix of latent proficiency variance and expected error variance. Moreover, the sum of $\sigma^2(\theta)$ and $\mathscr{E}\sigma^2(\hat{\theta}|\theta)$ in the denominator is not equal to $\sigma^2(\hat{\theta})$, completely eliminating bias-related terms from the formula. Although it may be regarded as a more sensible coefficient than $R(\hat{\theta})_{VR}$, it has drawbacks because it engages the latent trait definition of proficiency, and the variance of actual proficiency estimates does not play any role.

It is interesting to note that $\sigma^2(\varepsilon)$ is never used as error variance for any of the variance-ratio reliability coefficients. According to the ANOVA identity, the overall error variance $\sigma^2(\varepsilon)$ can be formulated as (Lord, 1983):

$$\sigma^2(\varepsilon) = \sigma^2[\mathscr{E}(\varepsilon|\theta)] + \mathscr{E}\sigma^2(\varepsilon|\theta) = \sigma^2(BIAS) + \mathscr{E}\sigma^2(\hat{\theta}|\theta), \tag{37}$$

because $\mathscr{E}(\varepsilon|\theta) = \mathscr{E}(\hat{\theta} - \theta|\theta) = BIAS$ and $\sigma^2(\hat{\theta}|\theta) = \sigma^2(\varepsilon|\theta)$ for a single fixed test taker. Obviously, $\sigma^2(\varepsilon)$ includes both random error variance and variance attributable to bias, distinguishing it from the expected error variance $\mathscr{E}\sigma^2(\hat{\theta}|\theta)$ and the error variance defined in Equation 36.

Now, let us consider the parallel-forms coefficient, $R(\hat{\theta})_{PF} = \rho(\hat{\theta}, \hat{\theta}')$. This coefficient is, by definition, the Pearson correlation for a population between proficiency estimates obtained from two parallel forms. In cases where double testing is impractical, Lord (1983) derived a formula for $R(\hat{\theta})_{PF}$ that is expressed in terms of statistical quantities obtainable from a single administration of a test under the assumption of strictly parallel forms. Note that Samejima (1994) derived the same result as Lord (1983); however, she used the traditional assumptions of CTT, which may not hold exactly for $\hat{\theta}$. The parallel-forms reliability coefficient, $R(\hat{\theta})_{PF}$, is defined as

$$R(\hat{\theta})_{PF} \equiv \rho(\hat{\theta}, \hat{\theta}') = \frac{\sigma(\hat{\theta}, \hat{\theta}')}{\sigma(\hat{\theta})\sigma(\hat{\theta}')} = \frac{\sigma(\hat{\theta}, \hat{\theta}')}{\sigma^2(\hat{\theta})}, \tag{38}$$

where $\sigma(\hat{\theta}) = \sigma(\hat{\theta}')$ for strictly parallel forms. It can be shown that the asymptotically unbiased estimator of $\sigma(\hat{\theta}, \hat{\theta}')$ is

$$\sigma(\hat{\theta}, \hat{\theta}') = \sigma^2(\hat{\theta}) - \mathscr{E}\sigma^2(\hat{\theta}|\theta). \tag{39}$$

Therefore, the parallel-forms reliability coefficient of $\hat{\theta}$ becomes equal to $R(\hat{\theta})_{EVR}$:

$$\begin{aligned} R(\hat{\theta})_{PF} \equiv \rho(\hat{\theta},\hat{\theta}') &= \frac{\sigma(\hat{\theta},\hat{\theta}')}{\sigma^2(\hat{\theta})} = \frac{\sigma^2(\hat{\theta}) - \mathscr{E}\sigma^2(\hat{\theta}|\theta)}{\sigma^2(\hat{\theta})} \\ &= 1 - \frac{\mathscr{E}\sigma^2(\hat{\theta}|\theta)}{\sigma^2(\hat{\theta})} \equiv R(\hat{\theta})_{EVR}. \end{aligned} \tag{40}$$

The last coefficient to discuss is $R(\hat{\theta})_{SC}$, the squared-correlation reliability coefficient. S. Kim (2012) demonstrated that $R(\hat{\theta})_{SC}$ based on the linear regression of $\hat{\theta}$ on θ differs from $R(\hat{\theta})_{PF}$. Coefficient $R(\hat{\theta})_{SC}$ is defined as

$$R(\hat{\theta})_{SC} \equiv \rho^2(\hat{\theta},\theta) = \frac{[\sigma(\hat{\theta},\theta)]^2}{\sigma^2(\hat{\theta})\sigma^2(\theta)}, \tag{41}$$

where $\sigma(\hat{\theta},\theta) = \sigma[\mathscr{E}(\hat{\theta}|\theta),\theta] + \mathscr{E}\sigma(\hat{\theta},\theta|\theta) = \sigma^2(\theta) + \sigma(\theta, BIAS)$. Let $L(\varepsilon|\theta)$ denote the linear regression of ε on θ with the regression coefficient of $\beta(\varepsilon|\theta) = \sigma(\varepsilon,\theta)/\sigma^2(\theta) = \sigma(\theta, BIAS)/\sigma^2(\theta)$. It follows that

$$R(\hat{\theta})_{SC} = \frac{\sigma^2(\theta)}{\sigma^2(\hat{\theta})}[1 + \beta(\varepsilon|\theta)]^2 = R(\hat{\theta})_{VR} \times [1 + \beta(\varepsilon|\theta)]^2. \tag{42}$$

Because $\mathscr{E}(\varepsilon \mid \theta) = \mathscr{E}(\hat{\theta} - \theta \mid \theta) = BIAS \neq 0$, the slope of $L(\varepsilon|\theta)$, $\beta(\varepsilon \mid \theta)$, may not be zero. Then, it is clear from Equation 42 that $R(\hat{\theta})_{SC}$ and $R(\hat{\theta})_{VR}$ differ by a factor of $\left(1 + \beta_{\varepsilon|\theta}\right)^2$. If $\sigma(BIAS,\theta)$ is positive, which usually is the case, $R(\hat{\theta})_{SC}$ will be larger than $R(\hat{\theta})_{VR}$. One way to estimate $R(\hat{\theta})_{SC}$ would be to use a mathematical expression for bias, which then can be used to estimate $\beta(\varepsilon|\theta)$. For example, Lord (1983) provided the bias function for $\hat{\theta}$ of θ under the three-parameter logistic model.

As evident from the formula itself, $R(\hat{\theta})_{SC}$ measures the extent to which proficiency estimates are associated with *latent* proficiencies. We can consider a similar (squared) correlation coefficient that involves *expected* proficiencies, denoted $\rho^2(\hat{\theta}, \mathscr{E}\hat{\theta})$. It is arguably more correct to refer to $\rho^2(\hat{\theta}, \mathscr{E}\hat{\theta})$ as a reliability coefficient, while $R(\hat{\theta})_{SC}$ can be seen as a squared validity coefficient. The key distinction lies in the fact that the two coefficients use definitions of true proficiency that differ by a constant *BIAS* for each test taker. If the bias is negligible, the distinction between the two will be trivial.

In many practical applications, $\hat{\theta}$ is often treated as an unbiased estimator, and the variance of the bias and its covariance with θ are neglected in estimating reliability. When the bias is negligibly small, such as for a long test, it may be argued that the expected variance terms are close to the variances associated with latent proficiencies: that is, $\sigma^2(\hat{\theta}) \cong \sigma^2(\theta) + \sigma^2(\varepsilon|\theta) \cong \sigma^2(\theta) + \mathscr{E}\sigma^2(\hat{\theta}|\theta) \cong \sigma^2[\mathscr{E}(\hat{\theta}|\theta)] + \mathscr{E}\sigma^2(\hat{\theta}|\theta)$.

Table 5.2 Variance-Ratio Reliability Coefficients for the Maximum Likelihood Estimator

Coefficient	Variance of Proficiency Estimates	True Proficiency Variance	Error Variance
$R(\widehat{\theta})_{VR}$	$\sigma^2(\widehat{\theta})$	$\sigma^2(\theta)$	$\mathscr{E}\sigma^2(\widehat{\theta}\vert\theta) + \sigma^2(BIAS)$ $+2\sigma(\theta, BIAS)$
$R(\widehat{\theta})_{EVR} = R(\widehat{\theta})_{PF}$	$\sigma^2(\widehat{\theta})$	$\sigma^2[\mathscr{E}(\widehat{\theta}\vert\theta)] = \sigma^2(\theta) +$ $\sigma^2(BIAS) + 2\sigma(\theta, BIAS)$	$\mathscr{E}\sigma^2(\widehat{\theta}\vert\theta)$
$R(\widehat{\theta})_{MVR}$	$\sigma^2(\theta) + \mathscr{E}\sigma^2(\widehat{\theta}\vert\theta)$	$\sigma^2(\theta)$	$\mathscr{E}\sigma^2(\widehat{\theta}\vert\theta)$

Consequently, the three variance-ratio reliability coefficients will yield similar results: $R(\widehat{\theta})_{VR} \cong R(\widehat{\theta})_{EVR} \cong R(\widehat{\theta})_{MVR}$. The parallel-forms reliability coefficient, $R(\widehat{\theta})_{PF}$, has been proven to be equal to $R(\widehat{\theta})_{EVR}$. Coefficient $R(\widehat{\theta})_{SC}$ is limited to the linear regression assumption, and from the well-known statistical properties, $R(\widehat{\theta})_{EVR}$ is generally larger than $R(\widehat{\theta})_{SC}$. The two coefficients, $R(\widehat{\theta})_{SC}$ and $R(\widehat{\theta})_{EVR}$, will be identical only if bias is zero.

As highlighted by previous researchers (e.g., S. Kim, 2012; Lord, 1983), these coefficients are *not* estimating the same parameter and thus are not interchangeable. In general, the following inequalities are likely to hold: $R(\widehat{\theta})_{PF} = R(\widehat{\theta})_{EVR} \geq R(\widehat{\theta})_{SC} \neq R(\widehat{\theta})_{MVR} \geq R(\widehat{\theta})_{VR}$. The fundamental cause of these differences is the bias in $\widehat{\theta}$, and the extent to which these coefficients yield different results largely depends on the amount of bias. If it is suspected that bias is not negligible, the use of $R(\widehat{\theta})_{EVR}$ would be advisable. Table 5.2 summarizes the variance terms involved in each of the variance-ratio reliability coefficients.

The impact of bias in reliability coefficients can be alleviated by using modified ML estimators that adjust for the bias of $\widehat{\theta}$. For example, Lord (1983) derived a bias-correction function for $\widehat{\theta}$ under the three-parameter logistic model. Samejima (1993a, 1993b) expanded Lord's (1983) work to develop a bias function for any discrete item responses. Warm's (1989) weighted likelihood estimator is also known to reduce the bias of $\widehat{\theta}$. However, in general, there is a dearth of literature focused on the effect of bias in reliability for various IRT models, including polytomous and multidimensional models. So, caution needs to be exercised when deciding which formula to use in practice.

Bayesian Proficiency Estimates

The Bayesian proficiency estimator of primary interest in this section is the EAP estimator denoted $\tilde{\theta}$. A crucial difference between the ML and EAP estimators in conceptualizing reliability lies in the use of different conditional distributions. The ML estimator focuses on the distribution of $f(\widehat{\theta}|\theta)$, while the EAP estimator is concerned with $f(\theta|\tilde{\theta})$. The variance of $f(\theta|\tilde{\theta})$ serves as the conditional error variance for EAP estimates, denoted $\sigma^2(\theta|\tilde{\theta})$. The commonly discussed and utilized reliability coefficients for $\tilde{\theta}$ include:

$$R(\tilde{\theta})_{SC} \equiv \rho^2(\theta,\tilde{\theta});$$

$$R(\tilde{\theta})_{VR} \equiv \frac{\sigma^2(\tilde{\theta})}{\sigma^2(\theta)} = \frac{\sigma^2[\mathscr{E}(\theta|\tilde{\theta})]}{\sigma^2(\theta)} = 1 - \frac{\mathscr{E}\sigma^2(\theta|\tilde{\theta})}{\sigma^2(\theta)} = \frac{\sigma^2(\tilde{\theta})}{\sigma^2(\tilde{\theta}) + \mathscr{E}\sigma^2(\theta|\tilde{\theta})}\text{; and}$$

$$R(\tilde{\theta})_{INF} \equiv 1 - \frac{\mathscr{E}[1/(I(\theta,\hat{\theta})+1)]}{\sigma^2(\theta)}.$$

The parallel-forms definition of reliability is *not* applicable to $\tilde{\theta}$ in the Bayesian framework because the obtained data are considered fixed. The squared-correlation definition of $R(\tilde{\theta})_{SC}$ resembles that of $R(\hat{\theta})_{SC}$ for the ML estimator, differing in that $R(\tilde{\theta})_{SC}$ is based on the linear regression of θ and $\tilde{\theta}$, $L(\theta|\tilde{\theta})$, while $R(\hat{\theta})_{SC}$ depends on $L(\hat{\theta}|\theta)$. Unlike the ML estimator, all variations of the variance-ratio coefficients associated with $R(\tilde{\theta})_{VR}$ are identical. Notably, the reversal of variances in true proficiencies and proficiency estimates in $R(\tilde{\theta})_{VR}$ compared to the ML estimator stems from the Bayesian framework's focus on predicting θ given $\tilde{\theta}$, as opposed to estimating $\hat{\theta}$ given θ under the ML framework. The last coefficient, $R(\tilde{\theta})_{INF}$, is proposed as an alternative reliability coefficient for $\tilde{\theta}$, approximating the relationship between the test information function (noted by the *INF* subscript) and the variance of the posterior distribution. As an approximation to $R(\tilde{\theta})_{VR}$, $R(\tilde{\theta})_{INF}$ offers the benefit of avoiding intricate computations of posterior variance while integrating the well-established concept of the test information function.

The EAP estimate for a test taker with item response data $\boldsymbol{u} = \{u_1,\ldots,u_n\}$, by definition, is the mean of the posterior distribution, which is given by

$$\tilde{\theta} = \mathscr{E}(\theta|\boldsymbol{u}) = \frac{\int_\theta \theta \prod_{i=1}^{n} \Pr(U_i = u_i|\theta) f(\theta) d\theta}{\int_\theta \prod_{i=1}^{n} \Pr(U_i = u_i|\theta) f(\theta) d\theta}, \tag{43}$$

where $\Pr(U_i = u_i|\theta)$ is the item response function for a given IRT model and $f(\theta)$ is the prior distribution of θ, which often is assumed to be the same for all test takers. The variance of the posterior distribution (i.e., error variance) for a test taker with $\boldsymbol{u}$ is (Bock & Mislevy, 1982):

$$\sigma^2(\theta|\boldsymbol{u}) = \frac{\int_\theta (\theta - \tilde{\theta})^2 \prod_{i=1}^{n} \Pr(U_i = u_i|\theta) f(\theta) d\theta}{\int_\theta \prod_{i=1}^{n} \Pr(U_i = u_i|\theta) f(\theta) d\theta}. \tag{44}$$

The integrals in Equations 43 and 44 are replaced by summations using a discrete proficiency distribution for computational purposes.

Of note, one distinct property of $\tilde{\theta}$ is its one-to-one correspondence with $\boldsymbol{u}$, meaning that test takers with the same response pattern will have the same $\tilde{\theta}$. This property suggests that the conditional posterior distribution of θ is identical using either $\boldsymbol{u}$ or $\tilde{\theta}$

as the conditioning variable. As such, $\tilde{\theta} = \mathscr{E}(\theta|\boldsymbol{u}) = \mathscr{E}(\theta|\tilde{\theta})$ and $\sigma^2(\theta|\boldsymbol{u}) = \sigma^2(\theta|\tilde{\theta})$. In the context of predicting θ given $\tilde{\theta}$, the ANOVA identity suggests that

$$\sigma^2(\theta) = \sigma^2[\mathscr{E}(\theta \mid \tilde{\theta})] + \mathscr{E}\sigma^2(\theta \mid \tilde{\theta}). \tag{45}$$

Based on this equality, the intraclass correlation coefficient is $\sigma^2[\mathscr{E}(\theta|\tilde{\theta})]/\sigma^2(\theta) = 1 - [\mathscr{E}\sigma^2(\theta|\tilde{\theta})/\sigma^2(\theta)]$.

One appealing feature of the Bayesian framework is that the distinction of expected variance terms is no longer necessary. The first term in the right-hand side of Equation 45, $\sigma^2[\mathscr{E}(\theta \mid \tilde{\theta})]$, is equal to $\sigma^2(\tilde{\theta})$ because $\mathscr{E}(\theta \mid \tilde{\theta}) = \tilde{\theta}$. The equality of $\sigma^2[\mathscr{E}(\theta \mid \tilde{\theta})] = \sigma^2(\tilde{\theta})$ is sufficient to render the first three variance-ratio definitions in $R(\tilde{\theta})_{VR}$ identical. It can be further shown that $\sigma^2(\varepsilon) = \sigma^2[\mathscr{E}(\varepsilon|\tilde{\theta})] + \mathscr{E}\sigma^2(\varepsilon|\tilde{\theta})$, similar to Equation 37, and because $\sigma^2[\mathscr{E}(\varepsilon|\tilde{\theta})] = 0$, $\sigma^2(\varepsilon) = \mathscr{E}\sigma^2(\varepsilon|\tilde{\theta}) = \mathscr{E}\sigma^2(\theta|\tilde{\theta})$. As a result,

$$\sigma^2(\theta) = \sigma^2(\tilde{\theta}) + \sigma^2(\varepsilon) = \sigma^2(\tilde{\theta}) + \mathscr{E}\sigma^2(\varepsilon \mid \tilde{\theta}). \tag{46}$$

It follows that various variance-ratio definitions ascribed to $R(\tilde{\theta})_{VR}$ are all equal to one another.

Similar to Equation 42 for the ML estimator, the EAP version of the squared-correlation coefficient is given by

$$R(\tilde{\theta})_{SC} \equiv \frac{\sigma^2[L(\theta|\tilde{\theta})]}{\sigma^2(\theta)} = \frac{\sigma^2(\tilde{\theta})}{\sigma^2(\theta)}. \tag{47}$$

Equation 47 clearly suggests that $R(\tilde{\theta})_{SC} = R(\tilde{\theta})_{VR}$. So, for the EAP estimator, $R(\tilde{\theta})_{SC}$ and various alternative definitions associated with $R(\tilde{\theta})_{VR}$ are all equal. This equality occurs as a result of the following one-to-one correspondence: $\mathscr{E}(\theta|\boldsymbol{u}) = \mathscr{E}(\theta|\tilde{\theta}) = \tilde{\theta}$. If the one-to-one correspondence is absent for an estimator, the equality may not hold. For example, S. Kim (2012) cautioned that another Bayes estimator, MAP, does not necessarily exhibit the one-to-one correspondence with the one- or two-parameter logistic IRT models.

The coefficient $R(\tilde{\theta})_{INF}$ incorporates the concept of information, typically associated with the ML estimator, to estimate the conditional and overall error variances. The link between the information for $\tilde{\theta}$ and $\sigma^2(\theta \mid \tilde{\theta})$ is established by the argument that the information provided by the prior distribution of θ is equivalent to adding an item to which all test takers in the population respond identically (Ferrando & Lorenzo-Seva, 2007; Thissen & Orlando, 2001). If the prior distribution of θ follows the standard normal distribution, as is commonly assumed, its contribution to the test information function is a constant value equal to 1 because $1/\sigma^2(\theta) = 1$. Consequently,

$$\sigma^2_{INF}(\theta \mid \tilde{\theta}) \cong \frac{1}{I(\theta,\hat{\theta}) + 1}, \tag{48}$$

where $I(\theta, \hat{\theta})$ is the traditional test information for a given θ. Furthermore, $\mathscr{E}\sigma^2(\theta \mid \tilde{\theta})$ in $R(\tilde{\theta})_{VR}$ can be replaced with $\sigma^2_{INF}(\theta|\tilde{\theta})$ computed by Equation 48 to derive an alternative expression for the reliability coefficient for $\tilde{\theta}$:

$$R(\tilde{\theta})_{INF} = 1 - \frac{\mathscr{E}[1/(I(\theta, \hat{\theta}) + 1)]}{\sigma^2(\theta)}, \tag{49}$$

which can be further reduced to $R(\tilde{\theta})_{INF} = 1 - \mathscr{E}[1/(I(\theta, \hat{\theta}) + 1)]$ because $\sigma^2(\theta)$ is equal to 1.

It is fair to say that coefficient $R(\tilde{\theta})_{INF}$ is computationally more convenient and capitalizes on the useful concept of information concerning its error variance. However, $R(\tilde{\theta})_{INF}$ would function merely as an approximation to the more defensible coefficient $R(\tilde{\theta})_{VR}$. This is because the algebraic similarity in aggregate statistics does not necessarily imply that the interpretation of error variance and its relation to other variance terms, especially in terms of conditional statistics, remains unchanged.

Summed Raw Scores and Transformed Scale Scores

IRT can serve as an underlying psychometric framework for estimating reliability for both summed raw scores, denoted X, and transformed scale scores, denoted $S = t(X)$. The general procedures suggested in the literature on this topic are largely consistent, except for variations in presentation details (e.g., S. Kim & Feldt, 2010; Kolen et al., 1996; Lord, 1980). As a function of a proficiency distribution and item parameters, the probability distributions for all three components of a measurement model (i.e., observed, true, and error scores) can be modeled and used subsequently to form reliability statistics.

A general model for representing marginal observed scores is given by

$$f(x) = \Pr(X = x) = \int_{\theta} \Pr(X = x|\theta) f(\theta) d\theta, \tag{50}$$

where $\Pr(X = x \mid \theta)$ is the conditional raw-score distribution and $f(\theta)$ is the distribution of θ. The conditional distribution, $f(x|\theta) = \Pr(X = x|\theta)$, captures the score variability for a test taker with θ over repeated testing using the strictly parallel forms or a fixed form. It can be efficiently computed using a recursive formula (e.g., Hanson, 1994; Lord & Wingersky, 1984; Thissen et al., 1995). Assuming a specific θ distribution (e.g., standard normal), the marginal observed score distribution, $f(x)$, is obtained by integrating the conditional distributions. The resulting $f(x)$ is often called the fitted or model-based observed score distribution, which lends itself to many applications in measurement such as test score equating, model fit examination, score smoothing, and quantifying reliability statistics.

For a test taker with θ, the expected value of $f(x|\theta)$ is the IRT analogue of true score in CTT, denoted

$$\tau_\theta = \mathscr{E}(X|\theta) = \sum x \Pr(X = x|\theta), \tag{51}$$

where the summation is taken over all possible values of X. The IRT true score τ_θ, a monotone nonlinear transformation of θ, is also referred to as the TCC. The variance of $\Pr(X = x|\theta)$ is the conditional error variance, computed by

$$\sigma^2(X|\theta) = \sum x^2 \Pr(X = x|\theta) - \tau_\theta^2. \tag{52}$$

The expected value of $\sigma^2(X|\theta)$ over all test takers in the population is the overall error variance, that is, $\mathscr{E}\sigma^2(X|\theta) = \int_\theta \sigma^2(X|\theta) f(\theta) d\theta$.

The ANOVA decomposition of the observed score variance is given by

$$\sigma^2(X) = \sigma^2[\mathscr{E}(X|\theta)] + \mathscr{E}\sigma^2(X|\theta) = \sigma^2(\tau_\theta) + \mathscr{E}\sigma^2(X|\theta). \tag{53}$$

Based on this ANOVA identity, the most widely used reliability coefficient is the following intraclass correlation:

$$R(X)_{IRT} = \frac{\sigma^2(\tau_\theta)}{\sigma^2(\tau_\theta) + \mathscr{E}\sigma^2(X|\theta)} = 1 - \frac{\mathscr{E}\sigma^2(X|\theta)}{\sigma^2(X)}. \tag{54}$$

S. Kim and Feldt (2010) referred to the first formula as the true score variance approach and the second formula as the observed score variance approach. Andersson and Xin (2018) called $R(X)_{IRT}$ the test reliability. The true score variance, $\sigma^2(\tau_\theta)$, in Equation 54 can be obtained using

$$\sigma^2(\tau_\theta) = \int_\theta \tau_\theta^2 f(\theta) d\theta - (\mathscr{E}\tau_\theta)^2, \tag{55}$$

where $\mathscr{E}\tau_\theta = \int_\theta \tau_\theta f(\theta) d\theta$. The second formula of $R(X)_{IRT}$ requires observed score variance, which is the variance of $f(x)$ in Equation 50. If the sample observed score variance, denoted $\hat{\sigma}^2(X)$, is used, the variance equality in Equation 53 is lost, and the two formulas of $R(X)_{IRT}$ are no longer exchangeable. The coefficient, $1 - [\mathscr{E}\sigma^2(X|\theta)/\hat{\sigma}^2(X)]$, might be best viewed as a sample estimate of $R(X)_{IRT}$, although there is no clear theoretical justification for doing so. The sample estimate may work well if the model fit is reasonably good and may sometimes be the only option when the model-based observed score variance cannot be computed easily (e.g., noninteger observed scores).

Note that there is a one-to-one correspondence between θ and τ_θ—that is, $\mathscr{E}(X \mid \theta) = \mathscr{E}[X \mid \tau_\theta] = \tau_\theta$. As a result, the squared-correlation coefficient $\rho^2(X,\tau)$ will be identical to $R(X)_{IRT}$.

Now let us consider scale scores, $S = t(X)$, that are transformed from raw scores. According to the well-known statistical property of correlation, a linear transformation of variables does not alter the value of correlation, assuming the direction of transformation for both variables is the same. This indicates that any linearly transformed scale scores will have the same reliability as that for raw scores. Thus, linear transformation is not discussed further in this section. By contrast, nonlinear transformation requires special treatment, as discussed next.

Suppose a conversion table contains all possible raw-score points, each corresponding to a scale-score point with a nonlinear relationship. The raw-to-scale score conversion table can be one-to-one or many-to-one and may contain integer or noninteger scale scores. Conceptually, over repeated measurements, the probability of a test taker receiving a certain scale score is the same as the probability of obtaining the corresponding raw score—that is, $\Pr[t(x)|\theta] = \Pr(X = x|\theta)$. The expected value of scale scores over replications for a test taker with θ is called *true scale score* for the test taker and is given by

$$\xi_\theta = \mathscr{E}(S|\theta) = \sum t(x)\Pr(X = x|\theta), \tag{56}$$

where the summation is over the entire range of raw scores. Likewise, the variance of scale scores for given θ is defined as

$$\sigma^2(S|\theta) = \sum [t(x)]^2 \Pr(X = x|\theta) - \xi_\theta^2. \tag{57}$$

The square root of Equation 57 is the *conditional scale-score SEM*. A reliability coefficient for scale scores can be expressed as

$$R(S)_{IRT} = \frac{\sigma^2(\xi_\theta)}{\sigma^2(\xi_\theta) + \mathscr{E}\sigma^2(S|\theta)} = 1 - \frac{\mathscr{E}\sigma^2(S|\theta)}{\sigma^2(S)}, \tag{58}$$

where $\mathscr{E}\sigma^2(S \mid \theta) = \int_\theta \sigma^2(S \mid \theta)g(\theta)d\theta$ and $\sigma^2(\xi_\theta)$ can be obtained similarly to Equation 55.

Composite Scores and Composite Scale Scores

A composite score, as defined earlier, is a weighted sum of scores from multiple constituent components. The potential score points and range of the composite score depend on the combinations of component scores and weights. In practical applications, composite scores are often rounded to the nearest integer, especially when noninteger weights are employed. However, rounding is not obligatory as long as all possible composite scores can be accurately identified and modeled.

The framework presented in this section closely aligns with the work of W. Lee et al. (2020), and similar developments can be found in Kolen et al. (2012) and Kolen and Lee (2011). W. Lee et al. (2020) considered three underlying IRT frameworks: unidimensional IRT (UIRT), simple-structure MIRT (SS-MIRT), and bifactor MIRT (BF-MIRT) models, each addressing potential multidimensionality in different ways. While the assumption of unidimensionality may be somewhat violated for composite scores, UIRT models are still commonly used. In the SS-MIRT model, each item is assumed to measure a single construct (i.e., component) and constructs across multiple components can be correlated (Kolen et al., 2012; W. Lee et al., 2020). The BF-MIRT model (Gibbons & Hedeker, 1992; Gibbons et al., 2007) assumes that all items in the test measure the same general construct, and each item measures an additional construct specific to the associated component. The general and specific constructs are typically uncorrelated. Further information about BF-MIRT can be found in Cai et al. (2011) and DeMars (2006, 2013).

The modeling of marginal composite scores follows a structure similar to Equation 50, with the use of a vector $\boldsymbol{\theta}$ as the conditioning variable:

$$f(z) = \Pr(Z = z) = \int_{\boldsymbol{\theta}} \Pr(Z = z|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{59}$$

The number of proficiency parameters in the vector $\boldsymbol{\theta}$ depends on the model—for UIRT models, $\boldsymbol{\theta} = \theta$; for the SS-MIRT model, $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$; and for the BF-MIRT model,

$\boldsymbol{\theta} = \{\theta_G, \theta_1, \ldots, \theta_M\}$, where θ_G indicates the general factor. Numerical integration can be carried out using a univariate standard normal distribution for UIRT or a multivariate normal density function for the MIRT models, with nonzero correlations between components for the SS-MIRT or zero correlations (typically) for the BF-MIRT.

Assuming conditional local independence, one way to compute the conditional composite score distribution, $f(z|\boldsymbol{\theta}) = \Pr(Z = z|\boldsymbol{\theta})$, is

$$\Pr(Z = z|\boldsymbol{\theta}) = \sum_{Z=w_1X_1+\cdots+w_mX_m} \prod_{m=1}^{M} \Pr\left(X_m = x_m|\boldsymbol{\theta}\right), \tag{60}$$

where the summation is taken over all possible combinations of weighted component scores that lead to the same composite score. The conditional distribution for each component can be computed using a recursion formula such as Lord and Wingersky's (1984) method. Note that not all elements of θ are used in computing the conditional distribution for each component. For example, for component m, only a single θ_m is involved for the SS-MIRT model, while both θ_G and θ_m are needed for the BF-MIRT model. The mean (i.e., true composite score) and variance of the conditional distribution, respectively, are

$$\zeta_{\boldsymbol{\theta}} = \mathscr{E}(Z|\boldsymbol{\theta}) = \sum z \Pr(Z = z|\boldsymbol{\theta}), \tag{61}$$

and

$$\sigma^2(Z|\boldsymbol{\theta}) = \sum z^2 \Pr(Z = z|\boldsymbol{\theta}) - \zeta_\theta^2. \tag{62}$$

A reliability coefficient then is given by

$$R(Z)_{IRT} = \frac{\sigma^2(\zeta_{\boldsymbol{\theta}})}{\sigma^2(\zeta_{\boldsymbol{\theta}}) + \mathscr{E}\sigma^2(Z|\boldsymbol{\theta})} = 1 - \frac{\mathscr{E}\sigma^2(Z|\boldsymbol{\theta})}{\sigma^2(Z)}. \tag{63}$$

Alternatively, $R(Z)_{IRT}$ can be expressed in the form of the general reliability-coefficient formula for linear composite scores (i.e., Equation 15), which involves error variances for individual components. That is,

$$R(Z)_{IRT^*} = 1 - \frac{\sum_{m=1}^{M} \mathscr{E}\sigma^2\left(X_m|\boldsymbol{\theta}\right)}{\sigma^2(Z)}, \tag{64}$$

where $\mathscr{E}\sigma^2\left(X_m|\boldsymbol{\theta}\right)$ is the overall error variance for each component and errors across components are assumed to be uncorrelated. Comparing $R(Z)_{IRT}$ and $R(Z)_{IRT*}$, it is evident that computing each $\mathscr{E}\sigma^2\left(X_m|\boldsymbol{\theta}\right)$ and adding them up is easier than computing $\mathscr{E}\sigma^2\left(Z|\boldsymbol{\theta}\right)$, which relies on the computationally cumbersome conditional composite score distribution, $f(z|\boldsymbol{\theta})$. However, $\sigma^2(Z)$ in the denominator of both coefficients is based on $f(z)$ in Equation 59, which, in turn, requires the use of $f(z|\boldsymbol{\theta})$. Therefore, bypassing $f(z|\boldsymbol{\theta})$ in the computational process is not a viable option. If the sample composite score variance, $\hat{\sigma}^2(Z)$, is used in place of the model-based one for $R(Z)_{IRT^*}$, some of the computational complexity will be alleviated. Doing so, however, will result in a reliability estimate that is best viewed as a sample estimate of $R(Z)_{IRT^*}$ (or $R(Z)_{IRT}$) and many desired features of the coherent psychometric framework will no longer be available. Yet, the computational simplicity might be very useful in practice.

Composite scores usually are further transformed nonlinearly to scale scores for reporting purposes. The results for composite scale scores, $S = t(Z)$, can be obtained in an analogous manner by noting that $\Pr[t(z)|\boldsymbol{\theta})] = \Pr(Z = z|\boldsymbol{\theta})$ and $\Pr[t(z)] = \Pr(Z = z)$. Replacing Z with $t(z)$ in Equations 61, 62, and 63 yields the true score, conditional error variance, and reliability coefficient all on the composite scale score metric.

Scale Scores Transformed From Proficiency Estimates

In situations where IRT proficiency scoring is used, the proficiency estimates derived from ML or EAP estimation methods are commonly converted to scale scores for reporting purposes. Here, we are mainly concerned with nonlinear transformations. Let $t(\theta)$ be a monotone increasing transformation function that converts θ to scale scores. It is assumed that the transformation is applied to both true and estimated proficiencies such that $S = t(\hat{\theta})$ and $\mathscr{E}(S|\theta) = \xi_\theta = t(\theta)$. The transformation function is further assumed to be continuous and differentiable at every level of θ. In cases where a mathematical expression for a theta-to-scale score transformation is not available, a practical solution might involve employing a smoothing technique to obtain a smooth continuous conversion relationship, such as a high-degree polynomial.

Let us begin with the ML estimator. According to Lord (1980, p. 67), the test information of observed scale score, S, for making inferences about θ, is defined as

$$I(\theta,S) = \frac{\left[\frac{\partial}{\partial\theta}\mathscr{E}(S|\theta)\right]^2}{\sigma^2(S|\theta)} = \frac{\left[t'(\theta)\right]^2}{\sigma^2(S|\theta)}, \tag{65}$$

where $t'(\theta)$ is the first derivative function of $t(\theta)$ and

$$\sigma^2(S|\theta) \cong \left[t'(\theta)\right]^2 \sigma^2(\hat{\theta}|\theta). \tag{66}$$

Equation 66 assumes that $\hat{\theta}$ is an asymptotically unbiased estimator of θ and is known as the delta method approximation (Kendall & Stuart, 1977), which will be detailed in the next section. Equations 65 and 66 suggest that $I(\theta,S) = I(\theta,\hat{\theta})$. Furthermore, the monotone increasing relationship between θ and $\xi_\theta \equiv \mathscr{E}(S|\theta)$ implies that $\sigma^2(S|\theta) = \sigma^2(S|\xi_\theta)$. The overall scale-score error variance can be obtained by integrating $\sigma^2(S \mid \boldsymbol{\theta})$ over the proficiency distribution as

$$\mathscr{E}\sigma^2(S \mid \theta) \cong \mathscr{E}\left(\frac{\left[t'(\theta)\right]^2}{I(\theta,S)}\right) = \mathscr{E}\left(\frac{\left[t'(\theta)\right]^2}{I(\theta,\hat{\theta})}\right). \tag{67}$$

The variance of true scale scores is

$$\sigma^2(\xi_\theta) = \int_\theta \xi_\theta^2 f(\theta)d\theta - (\mathscr{E}\xi_\theta)^2, \tag{68}$$

which can be evaluated using numerical integration. The scale-score analogues of $R(\widehat{\theta})_{VR}$, $R(\widehat{\theta})_{EVR}$, and $R(\widehat{\theta})_{MVR}$, respectively, are

$$R(S_{\hat{\theta}})_{VR} = \frac{\sigma^2(\xi_\theta)}{\sigma^2(S)};\ R(S_{\hat{\theta}})_{EVR} = 1 - \frac{\mathscr{E}\sigma^2(S|\theta)}{\sigma^2(S)};\ \text{and}$$

$$R(S_{\hat{\theta}})_{MVR} = \frac{\sigma^2(\xi_\theta)}{\sigma^2(\xi_\theta) + \mathscr{E}\sigma^2(S|\theta)}.$$

Similar to ML estimates, these three coefficients for scale scores do not necessarily yield the same result.

For the EAP estimator, the conditional error variance, $\sigma^2(\theta|\boldsymbol{u}) = \sigma^2(\theta|\tilde{\theta})$, is provided in Equation 44. The conditional error variance for scale scores can be found using the delta method approximation as

$$\sigma^2(\xi_\theta \mid \tilde{\theta}) \cong \left\{t'\,[\mathscr{E}(\theta|\tilde{\theta})]\right\}^2 \sigma^2(\theta|\tilde{\theta}) = \left[t'(\tilde{\theta})\right]^2 \sigma^2(\theta|\tilde{\theta}). \tag{69}$$

Note that, under the Bayesian framework, $\sigma^2(\xi_\theta \mid \tilde{\theta})$ is the variance of *true* scale scores conditional on $\tilde{\theta}$. It immediately follows that the overall scale-score error variance is $\mathscr{E}\sigma^2(\xi_\theta|\tilde{\theta}) \cong \mathscr{E}\left\{\left[t'(\tilde{\theta})\right]^2 \sigma^2(\theta|\tilde{\theta})\right\}$, where the expectation (i.e., average) is taken over all persons in the data. A reliability coefficient for scale scores takes a form similar to $R(\tilde{\theta})_{VR}$:

$$R(S_{\tilde{\theta}})_{VR} = \frac{\sigma^2(S)}{\sigma^2(\xi_\theta)} \cong 1 - \frac{\mathscr{E}\sigma^2(\xi_\theta|\tilde{\theta})}{\sigma^2(\xi_\theta)} \cong \frac{\sigma^2(S)}{\sigma^2(S) + \mathscr{E}\sigma^2(\xi|\tilde{\theta})}, \tag{70}$$

where the true scale-score variance, $\sigma^2(\xi_\theta)$, can be computed using Equation 68. Unlike the results for $\tilde{\theta}$, the three variations of $R(S_{\tilde{\theta}})_{VR}$ may not yield the same results because of the use of the delta method approximation and possible rounding of reported scale scores.

As discussed earlier, the Bayesian conditional error variance can also be approximated using the test information function as $\sigma^2_{INF}(\theta \mid \tilde{\theta}) \cong 1 / [I(\theta,\theta) + 1]$. In this approximation, the conditioning variable becomes true θ and a discrete set of quadrature points can be used instead of individual $\tilde{\theta}$'s. Then, the conditional scale-score variance can be estimated using the delta method, which gives $\sigma^2(\xi_\theta|\theta) \cong \left[t'(\theta)\right]^2 \sigma^2_{INF}(\theta|\tilde{\theta})$, and the overall error variance is obtained by integrating $\sigma^2(\xi_\theta|\theta)$ over the θ quadrature distribution. Finally, the scale-score version of $R(\tilde{\theta})_{INF}$ is given by $R(S_{\tilde{\theta}})_{INF} = 1 - \left[\mathscr{E}\sigma^2(\xi_\theta|\theta)/\sigma^2(\xi_\theta)\right]$.

ESTIMATORS OF CSEMs

Conceptually, SEM represents the standard deviation of test scores over repeated measurements for a person, assuming the measurement process does not alter their true score. In CTT, the SEM is commonly expressed as $\sigma(E) = \sigma(X)\sqrt{1-\rho(X,X')}$, referred to as the overall SEM. While this computation offers simplicity, an excessive focus on reliability coefficients may divert attention from actively considering whether the sources of error involved in the SEM closely align with the intended use and generalization of the test results. Additionally, applying the same SEM to all test takers when constructing confidence intervals overlooks variations in SEM magnitude as a function of true and observed scores. An arguably more defensible approach might involve starting with CSEMs, which can then be used to derive both the overall SEM [see Equation 6] and reliability coefficients, if needed.

Although direct estimation of CSEMs over many repeated measurements is practically unfeasible, collecting data from two replications may not be so unrealistic. This approach is similar to the data collection design used for estimating the test–retest and parallel-forms reliability coefficients. Using two sets of observed scores, X_1 and X_2, an estimate of the CSEM for a single test taker can be computed directly as

$$CSEM_R = \frac{|X_1 - X_2|}{\sqrt{2}}. \qquad (71)$$

Some benefits of Equation 71 were argued by Brennan (2001a), including that, by replicating the full-length measurement procedure twice, the investigator is forced to focus on the actual characteristics of replications to ensure that they are faithful reflections of the intended universe. In addition, Equation 71 provides a direct estimate of the CSEM without making any assumptions, applicable to various score types such as raw scores, scale scores, and composite scores. This is especially advantageous given that computing CSEMs for scale scores based on a single administration of a test can often be complicated (e.g., Kolen et al., 1992, 1996; W. Lee et al., 2000). Despite these potential benefits, Equation 71 is not widely used in practice, partly because of the questionable accuracy of estimation based on only two replications and the challenge of obtaining two full-length replications on adequate samples.

Similar to reliability coefficients, considerable effort has been invested in developing estimators for CSEMs based on a single test administration using various psychometric models and making assumptions that are often too strong or untestable. In theory, CSEM is the standard deviation of observed scores over repeated parallel (classically, randomly, etc.) measurements conditional on each person's true score: that is, $\sigma(X|\tau_p) = \sigma(E|\tau_p)$. (Note that the subscript p denotes a particular person but may be omitted for simplicity when context allows.) Estimators for CSEMs fall broadly into two categories: those that condition on observed scores and those that condition on true scores. The former category avoids assumptions about the true score distribution, employing each test taker's observed score as an estimator of their true score, $\hat{\tau}_p = x_p$, since true scores are unknown.

Lord (1984) and Woodruff (1990) noted, however, that considering CSEM as conditional on observed score is appropriate only when a test taker at a particular observed score is randomly selected. Nonetheless, the bias in estimated CSEMs using observed scores as a conditioning variable diminishes with increasing test reliability. Also, studies have found this bias to be practically minimal (Feldt et al., 1985; W. Lee et al., 2000; Woodruff et al., 2013).

The second type of procedures make certain assumptions to estimate the true score distribution and express CSEMs conditional on true score. Consequently, these procedures eliminate the bias issue and ambiguity of interpretation. However, if reporting the CSEM to each individual test taker is the goal, mapping observed scores to true scores used for CSEM computation becomes inevitable, which makes the practice essentially the same as using observed scores as a conditioning variable. Both types of procedures generally yield similar results, and no particular argument is made here for preferring one approach over the other. Rather, it is advised that the choice of method should depend on the characteristics of available data, reasonableness of assumptions underlying the psychometric model, and available computer resources.

Summed Raw Scores

Thorndike Method

Thorndike (1951) was one of the first investigators to examine CSEMs that vary as a function of observed scores based on split-half tests. Suppose the total test score X is split into two essentially tau-equivalent half test scores, $X_1 = (1/2)T + C_1 + E_1$ and $X_2 = (1/2)T + C_2 + E_2$, and $X = X_1 + X_2$. It can be shown that, under CTT assumptions, the variance of difference scores between half-test scores is equal to the error variance for the total test:

$$\sigma^2(X_1 - X_2) = \sigma^2(C_1 - C_2 + E_1 - E_2) = \sigma^2(E_1) + \sigma^2(E_2) = \sigma^2(E). \qquad (72)$$

This implies that, for any group of test takers with total observed score x, the standard deviation of difference scores between X_1 and X_2 provides an estimate of the CSEM for that group:

$$CSEM(X)_{TH} = \sigma(X_1 - X_2 \,|\, x). \qquad (73)$$

However, a limitation arises as CSEMs across the range of total scores can be erratic, particularly at extreme scores due to low test-taker frequencies. In practice, test takers can be grouped into short intervals based on their total scores, and the CSEM is computed for each interval (i.e., subgroup).

Mollenkopf Method

Mollenkopf (1949) developed an estimator for CSEMs based on split halves, often considered a refinement of Thorndike's method (Feldt & Qualls, 1996). Assuming tau equivalence of the two split-half tests, consider the following quantity for person p:

$$D_p = [(X_{p1} - X_{p2}) - (\bar{X}_1 - \bar{X}_2)]^2, \qquad (74)$$

where $\bar{X}_1$ and $\bar{X}_2$ are group means on the two half-tests. For any subgroup of N test takers, the average of D_p values, by definition, is the variance of half-test difference scores, which is equal to the error variance for the full-length test:

$$\bar{D} = \frac{1}{N}\sum_{p=1}^{N}[(X_{p1} - X_{p2}) - (\bar{X}_1 - \bar{X}_2)]^2 = \sigma^2(X_1 - X_2) = \sigma^2(E). \tag{75}$$

The square root of $\bar{D}$ for a subgroup of test takers with an observed score x is the CSEM, which is identical to Thorndike's CSEM. Mollenkopf (1949) goes further to suggest using a polynomial regression with D_p in Equation 74 as a dependent variable and the total score X as an independent variable:

$$\hat{D} = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_q X^q, \tag{76}$$

where the polynomial degree q is chosen by the investigator. If the regression model fits well, the resulting predicted values, $\hat{D}$, provide a better approximation of conditional means than $\bar{D}$. Once the polynomial regression coefficients are estimated with a chosen q, Mollenkopf's CSEM at observed score x is

$$CSEM(X)_{MO} = \sqrt{\hat{D}} = \sqrt{\beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_q x^q}. \tag{77}$$

Although fitting a polynomial regression is cumbersome, $CSEM(X)_{MO}$ has the desirable characteristic of providing smoother patterns of CSEMs along the observed score scale.

Binomial Error Model

A notable drawback of $CSEM(X)_{TH}$ and $CSEM(X)_{MO}$ is that they require splitting the full test into two parallel halves, and there are various possibilities for obtaining such splits. Lord's (1955, 1957) seminal work on the binomial error model provided a substantially different perspective. In this framework, a test form with n dichotomously scored items is viewed as a random sample from a universe of such items (i.e., each item has an equal probability of being selected). Each test taker has a true proportion-correct score, $\pi_p = \tau_p/n$, such that the probability of answering each undifferentiable item correctly is π_p for test taker p. Measurement errors are conceptualized by the variation in observed scores across an infinite number of test forms, each containing a different set of randomly selected n items.

For a test taker with π_p, the distribution of observed scores from repeated measurements follows a binomial distribution with the parameter π_p, and the error variance for the test taker is $\sigma^2(E|\pi_p) = \sigma^2(X|\pi_p) = n\pi_p(1-\pi_p)$. Lord's CSEM, replacing the unknown parameter π_p with the observed proportion-correct score $\bar{x}_p = x_p/n$, is given by

$$CSEM(X)_{BN} = \hat{\sigma}(E|\bar{x}_p) = \sqrt{\frac{x_p(n - x_p)}{n-1}}. \tag{78}$$

This formula can compute CSEMs for raw-score points ranging from 0 to n without needing test-taker data. It has a quadratic form, ensuring the resulting CSEMs exhibit a smooth inverted-U shape with zero values at observed scores of zero and n.

Criticism of Lord's CSEM involves its failure to differentiate items, treating them as having equal difficulties under the binomial error model. In response, Keats (1957) proposed a correction, introducing a fuzzy factor of $(1-KR20)/(1-KR21)$ to multiply the squared values of Lord's CSEMs. This adjustment is intended to bring the overall error variance, when averaged, into alignment with the error variance in KR20. From the perspective of CTT, the criticism and Keats's adjustment find support. However, from the viewpoints of GT, Lord's model possesses valuable features that were not clearly identifiable at the time. The notion of random sampling of items over replications diverges from CTT definitions of parallelism but aligns with the key characteristic of GT—that is, the randomly parallel forms assumption. Moreover, Lord's error variance can be shown to be identical to the conditional *absolute* error variance for a single-facet design in GT. Lord's estimator indeed involves absolute error, a type of error not present in CTT. Therefore, Lord's binomial error model is sometimes considered a bridge between CTT and GT (Brennan, 2010).

Multinomial Error Model

Application of Lord's binomial error model is limited to items with binary score categories (i.e., right or wrong). W. Lee (2007) expanded Lord's model to cases where each item in an n-item test is scored polytomously with k possible score points, $a_1, a_2, \ldots, a_k$. A set of n items in a test is viewed as a random sample from an undifferentiated universe of such items. For notational simplicity, the derivation is presented for a single test taker without the subscript p. Let $\boldsymbol{\eta} = \{\eta_1, \eta_2, \ldots, \eta_k\}$ indicate the true category-proportion scores for a test taker; namely, η_1 is the proportion of items in the universe for which the test taker would get a score of a_1, η_2 for a_2, and so on. The sum of the η values is equal to 1. Let $Y_1, Y_2, \ldots, Y_k$ be random variables representing the observed numbers of items scored $a_1, a_2, \ldots, a_k$, respectively, such that $Y_1 + Y_2 + \ldots + Y_k = n$. It follows that the total raw score X is given as $X = a_1 Y_1 + a_2 Y_2 + \ldots + a_k Y_k$.

Due to the random sampling assumption, the category counts $Y_1, Y_2, \ldots, Y_k$ will follow a multinomial distribution:

$$f(y_1, y_2, \ldots, y_k \mid \boldsymbol{\eta}) = \frac{n!}{y_1! y_2! \cdots y_k!} \eta_1^{y_1} \eta_2^{y_2} \cdots \eta_k^{y_k}. \tag{79}$$

The conditional distribution of X can then be obtained as

$$f(x \mid \boldsymbol{\eta}) = \Pr(X = x \mid \boldsymbol{\eta}) = \sum_{x = a_1 y_1 + \cdots + a_k y_k} f(y_1, y_2, \ldots, y_k \mid \boldsymbol{\eta}), \tag{80}$$

where the summation is taken over all weighted combinations of $a_1 y_1 + \ldots + a_k y_k$ that sum to x.

Substituting $\hat{\eta}_i = y_i/n$ for η_i and applying a bias-correction factor, $\sqrt{n/(n-1)}$, an estimator of the CSEM for a test taker with observed category counts of $y_1, \ldots, y_k$ is given by

$$CSEM(X)_{MN} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{k} a_i^2 y_i (n-y_i) - 2\sum\sum_{i<j} a_i a_j y_i y_j\right]}. \tag{81}$$

Of note, using the above formula results in different estimated CSEMs for test takers with the same total raw score but varying configurations of category scores. W. Lee (2007) observed that multinomial CSEMs often display a vertically scattered umbrella pattern when plotted against the total raw score. When item scoring is binary, the multinomial estimates become identical to Lord's estimates.

Univariate GT

A general approach to estimating CSEMs under the GT framework has been developed (Brennan, 1998, 2001c). In GT, a distinction between absolute error and relative error is made even for the CSEMs. Estimating absolute CSEMs is relatively straightforward, while computing relative CSEMs is usually complicated although a simplified formula is available under certain assumptions. In GT, analyses commonly use the *mean* score metric rather than the total score metric, and results expressed in one metric can be easily transformed into the other. As discussed earlier, absolute error for a single person p is defined as $\Delta_p = \bar{X}_p - \mu_p$, the difference between the person's observed mean score and universe score. The variance of Δ_p over randomly parallel instances of a measurement procedure is the absolute error variance for the person. In general, for any random effects design, the absolute CSEM for person p in the *mean* score metric is

$$CSEM(\bar{X})_{GTabs} = \sigma^2(\Delta_p). \tag{82}$$

For example, if the D-study design under consideration is $p \times (I:H)$, the design for an *individual person* is $I:H$. The variance components $\sigma^2(I:H)_p$ and $\sigma^2(H)_p$ are computed using data for each individual person, highlighted by the subscript p below each variance component. Then, the absolute CSEM for person p in the *mean* score metric is $\sigma(\Delta_p) = \sqrt{\sigma^2(h)_p/n'_h + \sigma^2(i:h)_p/n'_i n'_h}$. Multiplying this result by $n'_i n'_h$ gives $\sigma(\Delta_p)$ in the total score metric.

Estimating *relative* CSEMs in GT is much more complicated, largely because of some nonzero covariance terms that are difficult to estimate (see Jarjoura, 1986). For practical purposes, Brennan (2001c) offered an approximate estimator for any random effects designs in the *mean* score metric as

$$CSEM(\bar{X})_{GTrel} = \sigma(\delta_p) \cong \sqrt{\sigma^2(\Delta_p) - \left[\sigma^2(\Delta) - \sigma^2(\delta)\right]}. \tag{83}$$

Estimating the relative CSEM using Equation 83 is a two-step process: (a) computing the absolute CSEM using Equation 82 and (b) making an adjustment to it by the difference between the absolute and relative *overall* error variances. Clearly, the adjustment to $\hat{\sigma}^2\left(\Delta_p\right)$ is a constant for all persons.

UIRT Proficiency Estimates

In the context of ML estimation, the item and test information functions are frequently used as a measure of conditional accuracy of estimation. If item parameters are known, the information functions do not depend on the group of test takers being tested. The conditional variance of ML estimates is inversely related to the test information function as $\sigma^2(\hat{\theta}\,|\,\theta) = 1/I(\theta, \hat{\theta})$ (Lord, 1980). Thus, $CSEM(\hat{\theta}) = \sqrt{1/I(\theta,\ \hat{\theta})}$. The concept of a test information function and its inverse to define error variance holds numerous convenient properties. A limitation, however, is that there is no general conceptual framework or statistical procedure for incorporating multiple sources of error in the estimation of error variance. An exception is the work by Bock et al. (2002), who proposed an approach to incorporating multiple ratings in IRT aiming for correcting for the bias in the standard error of proficiency estimates. Nevertheless, this approach is considered somewhat ad hoc, lacking a robust substantiation of the notion of replications (e.g., distinction between random and fixed facets) (Brennan, 2001a). Their approach certainly deserves further research.

Error variance in the Bayesian framework is defined as the variance of the posterior distribution of θ for a test taker with item response data $\boldsymbol{u}$, and its standard deviation is the CSEM for the test taker. Since there is a one-to-one correspondence between the EAP estimator $\tilde{\theta}$ and $\boldsymbol{u}$, test takers with identical item responses will share the same CSEM. The variance form of CSEM for $\tilde{\theta}$ was presented in Equation 44. Under this framework, the notion of replications plays no explicit role in conceptualizing the variability of θ's in the posterior distribution. To explicitly incorporate errors due to replications in estimating CSEMs and other reliability statistics, one approach is to apply the estimation procedure to at least two independent administrations of the test on the same group of test takers, preferably using different sets of items from the same domain. While this empirical approach is more feasible with computerized adaptive testing, a comprehensive theoretical framework has not been developed to seamlessly integrate the Bayesian definition of error based on the posterior distribution with one fixed set of data and the errors from actual replications involving more than one data set.

One well-known property of IRT proficiency estimates is that CSEMs tend to be smallest in the middle range of the proficiency scale and increase at both extremes. This phenomenon is not inherent to IRT but is a direct consequence of using the θ scale. As argued by Lord (1980), there is no unique virtue of the θ scale for measuring proficiency, and in general, there is no obvious reason to prefer θ over any other monotonic transformation of it. If θ is transformed to the (true) raw-score metric, the resulting CSEMs will exhibit a reversed pattern compared to that for the θ CSEMs. Nonlinear transformations from one metric to another not only change the magnitude of CSEMs but also alter their general pattern along the score scale, as discussed next.

Scale Scores

The current *Standards* (AERA et al., 2014) and its predecessors recommend (a) reporting standard errors at different score levels (i.e., conditional) and (b) expressing them in units of a reported score scale. Most reliability statistics, including CSEMs, rely on the specific score scale used for reporting. Two main approaches exist for estimating CSEMs for nonlinearly transformed scale scores. The first utilizes the delta method, a general statistical technique for estimating the variance of a statistic that is a function of another statistic with a known variance (Kendall & Stuart, 1977). The second approach involves computing the conditional distribution for scale scores under assumed psychometric models.

The Delta Method Approximation

To demonstrate how scale transformation affects CSEMs, we begin with the discussion of linear transformation. Let $S = t(X)$ be a linear transformation function of scores X to scale scores S such that $S = B + A(X)$, where A and B are the slope and intercept of the linear function, respectively. It follows that $\sigma^2(S) = A^2\sigma^2(X)$. This relationship also holds for conditional observed score (i.e., error) variances. That is, denoting E_X and E_S as error scores on the metrics of X and S, respectively, $\sigma^2(E_S|\tau) = A^2\sigma^2(E_X|\tau)$. This suggests that the scale-score CSEM can be expressed by multiplying the CSEM for scores X by the slope of the transformation function. Thus, for linear transformations, computations are straightforward if CSEM estimates for the original score scale, X, are available.

When dealing with nonlinear transformations, where the slope of the transformation is not constant across score levels, complexities arise. The delta method provides a useful approximation to address this issue. Consider a continuous, differentiable function, $S = t(X)$, where X can represent raw scores, IRT proficiency estimates, or composite scores. The variance of S can be approximated using the delta method as

$$\sigma^2(S) = \left(\frac{dt}{dX}(\mathscr{E}X)\right)^2 \sigma^2(X), \tag{84}$$

where $dt/dX = t'(X)$ is the first derivative function of $t(X)$ with respect to X, which is evaluated at the mean of X. In essence, the delta method approximates the variance of scale scores as a function of both the variance of X and the first derivative (i.e., slope) of the transformation function.

In real-world applications, $S = t(X)$ is rarely a continuous function with a concrete mathematical expression. More frequently, a conversion table contains a discrete set of values for X and corresponding scale-score points, usually rounded to integers. To attain a differentiable mathematical function, Feldt and Qualls (1998) proposed fitting a high-degree polynomial function to the discrete conversion table as

$$t(X) \cong \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_q x^q, \tag{85}$$

where the polynomial degree q can be determined based on visual inspection of the fit, incremental R-square statistics, or other relevant criteria. The first derivative of the polynomial function at $X = x$ is

$$t'(x) \cong \beta_1 + 2\beta_2 x + \cdots + q\beta_q x^{q-1}. \tag{86}$$

Once the first derivative function is computed using Equation 86, the scale-score CSEM can be approximated using any CSEM estimators for X. In general, the scale-score CSEM is estimated using this polynomial procedure as

$$CSEM(S)_{PN} = \sigma(E_S|\tau) = \sigma(S|\tau) \cong t'(\mathscr{E}X|\tau)\sigma(X|\tau). \tag{87}$$

For example, consider a nonlinear transformation of IRT θ to a certain type of scale scores with scoring performed using the EAP estimator. The first step is to fit a polynomial regression on the conversion relationship and find the first derivative function as in Equation 86. Then, an approximate value of the scale-score CSEM for a test taker with $\tilde{\theta}$ is $\sigma(S|\tilde{\theta}) \cong t'(\tilde{\theta})\sigma(\theta|\tilde{\theta})$ because $E(\theta|\tilde{\theta}) = \tilde{\theta}$. The polynomial procedure, as an approximation to the delta method, is versatile because it can be combined with any CSEM estimates for the original untransformed scores. However, a limitation is the somewhat subjective decision on the degree of the polynomial, and the estimated conditional slope based on the fitted function is still an approximation. The subsequent estimation procedures discussed do not necessitate such a slope-estimation process and utilize the entire distribution for scale scores.

Strong True Score Models

Kolen et al. (1992) extended Lord's (1965) strong true score theory to estimating scale-score CSEMs for tests scored number correct. Letting π denote the proportion-correct true score, the general expression for strong true score theory is given by

$$f(x) = \Pr(X = x) = \int_\pi \Pr(X = x|\pi) f(\pi) d\pi, \tag{88}$$

where the conditional distribution, $f(x|\pi) = \Pr(X = x|\pi)$, can be modeled using either a binomial or a two-term approximation to the compound binomial distribution; the marginal true score distribution, $f(\pi)$, is assumed to follow either a two-parameter or a four-parameter beta distribution; and $f(x)$ is the resulting marginal observed score distribution. Kolen et al. (1992) considered the most complex four-parameter beta compound binomial model.

The CSEM on the raw-score metric is expressed as

$$CSEM(X)_{ST} = \sigma(X|\pi) = \sqrt{\sum_{x=0}^{n} x^2 \Pr(X = x|\pi) - \left[\sum_{x=0}^{n} x \Pr(X = x|\pi)\right]^2}. \tag{89}$$

For a given discrete raw-to-scale score conversion table, the probability for earning a particular raw score is identical to the probability for earning the corresponding scale score—that is, $f[t(x)|\pi] = f(x|\pi)$. This allows replacing X with $S = t(X)$ in Equation 89 to obtain the scale-score CSEM:

$$CSEM(S)_{ST} = \sigma(S|\pi) = \sqrt{\sum_{x=0}^{n}[t(x)]^2 \Pr(X = x \mid \pi) - \left[\sum_{x=0}^{n} t(x)\Pr(X=x|\pi)\right]^2}. \quad (90)$$

Equation 90 provides unambiguous estimates of scale-score CSEMs no matter whether a conversion table is one-to-one or many-to-one (i.e., multiple raw-score points convert to a single scale-score point). If desired, the conditional scale-score distribution based on a many-to-one conversion table can be attained by summing the probabilities associated with all raw scores converting to a particular scale-score point. Doing this for all unique scale-score values will give a distribution for scale scores.

Binomial Procedure

For a test with dichotomous items, Brennan and Lee (1999) derived results for scale-score CSEMs, seen as a scale-score analogue of Lord's CSEM. The assumptions underlying this approach align with Lord's binomial error model. Since the binomial error model does not assume a specific true score distribution, the CSEM is estimated for each individual or uses observed score as a conditioning variable.

The conditional observed score distribution for an individual with a proportion-correct observed score $\bar{x} = \hat{\pi}$ follows a binomial distribution as

$$\Pr(X=x|\hat{\pi}) = \binom{n}{x}\hat{\pi}^{x}(1-\hat{\pi})^{n-x}. \quad (91)$$

Lord's CSEM is the square root of the unbiased estimate of variance of the binomial distribution, as shown in Equation 78: $\sqrt{n/(n-1)}\sqrt{[x(n-x)]/n}$, where $\sqrt{n/(n-1)}$ is a bias-correction factor. Lord's CSEM formula can be re-expressed as

$$\sigma(X|\hat{\pi}) = \sqrt{\frac{n}{n-1}}\sqrt{\sum_{x=0}^{n}x^2 \Pr(X=x|\hat{\pi}) - \left[\sum_{x=0}^{n}x\Pr(X=x|\hat{\pi})\right]^2}, \quad (92)$$

where the squared term inside the square root is the mean of the conditional distribution. By replacing raw score x with $t(x)$, the scale-score version of Lord's CSEM is

$$CSEM(S)_{BN} = \sqrt{\frac{n}{n-1}}\sqrt{\sum_{x=0}^{n}[t(x)]^2 \Pr(X=x|\hat{\pi}) - \left[\sum_{x=0}^{n}t(x)\Pr(X=x|\hat{\pi})\right]^2}. \quad (93)$$

Brennan and Lee (1999) discovered that the overall pattern and magnitude of scale-score CSEMs estimated based on the binomial procedure closely aligned with results from the polynomial-delta method.

Other Procedures

The expression for scale-score CSEMs in UIRT can be obtained by the square root of Equation 57. The derivation of scale-score CSEMs for the multinomial error model is outlined by W. Lee (2007), which is virtually the same as the process used for the binomial procedure. It involves using the fact that $f[t(x)|\eta] = f(x|\eta)$ (see Equation 80) and computing the standard deviation of the scale-score distribution.

Composite Scores

Compound Binomial and Compound Multinomial Error Models

The binomial error model has been extended to cases in which a test consists of multiple strata or components, known as the compound binomial model (Feldt, 1984). Let the number of items in each of M components be $n_1, n_2, \ldots, n_M$ and $X_1, X_2, \ldots, X_M$ represent the component scores. In this model, it is assumed that each item belongs to one of M distinct universes of items, errors follow a binomial distribution within each component, and errors are uncorrelated across components. Due to the uncorrelated-error assumption, the composite error variance is a simple weighted sum of Lord's (1957) error variances for the components. For a test taker with observed component scores $x_1, x_2, \ldots, x_M$, the estimated CSEM (often called Feldt's SEM) is

$$CSEM(Z)_{CB} = \sqrt{\sum_{m=1}^{M} w_m^2 \left[\frac{x_m(n_m - x_m)}{n_m - 1}\right]}. \tag{94}$$

Similarly, the extension of the multinomial error model to composite scores is provided by W. Lee (2007). The composite CSEM is the square root of a weighted sum of multinomial error variances for the various components, each of which is estimated using Equation 81. A constraint for the compound multinomial model is that the number of score categories must be the same for all items within a component.

Multivariate GT

As an extension of univariate GT, multivariate GT provides a framework for estimating CSEMs for composite scores. From the multivariate GT perspective, the foregoing treatment of composite scores is nothing more than a multivariate $p^\bullet \times i^\circ$ design. Let us define the composite score in the *mean* score metric as $\bar{Z} = \sum_{m=1}^{M} w_m \bar{X}_m$. Under the multivariate $p^\bullet \times i^\circ$ design, the within-person design for each person is simply i°; that is, items are nested within a set of fixed content categories. In such a case, the absolute CSEM for a person is the square root of a weighted sum of squared absolute CSEMs over the content categories: $\sigma(\Delta_{p\bar{Z}}) = \sqrt{\sum_m w_m^2 \sigma_m^2(\Delta_p)} = \sqrt{\sum_m w_m^2 \sigma_m^2(i)/n'_{im}}$. This result, when transformed to the total score metric, will be identical to the result based on the compound binomial model presented in Equation 94. Similarly, the relative composite score CSEM for a person is $\sigma(\delta_{p\bar{Z}}) = \sqrt{\sum_m w_m^2 \sigma_m^2(\delta_p)}$, where the relative CSEM for each category is computed using Equation 83.

The application of multivariate GT in the estimation of composite score CSEMs is not limited to the $p^{\bullet} \times i^{\circ}$ design and can be further extended to many multivariate designs that other competing approaches may not be able to handle properly. For example, consider an assessment that consists of two components, speaking and listening, each consisting of different performance tasks (i), and test takers' responses to both types of tasks are evaluated by the same set of human raters (r). Here, the two content components act as the fixed multivariate variable and the multivariate design is $p^{\bullet} \times i^{\circ} \times r^{\bullet}$. The within-person design is $i^{\circ} \times r^{\bullet}$ and the absolute error variance–covariance matrix is $\begin{bmatrix} \sigma_1^2(\Delta_p) & \sigma_{12}(\Delta_p) \\ \sigma_{12}(\Delta_p) & \sigma_2^2(\Delta_p) \end{bmatrix}$, where the subscripts 1 and 2 indicate the two components, and the variances and covariances are derived from the variance–covariance component matrices of $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Sigma}_r$, and $\boldsymbol{\Sigma}_{ir}$. Note that errors are allowed to be correlated here, which is unique to multivariate GT compared to other approaches that often assume uncorrelated errors.

A general formula for computing absolute error composite score CSEMs is given by

$$CSEM(\bar{Z})_{MGTabs} = \sigma_{\bar{Z}}(\Delta_p) = \sqrt{\sum_m w_m^2 \sigma_m^2(\Delta_p) + \sum_{m \neq m'}\sum w_m w_{m'} \sigma_{mm'}(\Delta_p)}. \quad (95)$$

Similar to Equation 83 for the univariate case, the relative error composite score CSEMs can be approximated as

$$CSEM(\bar{Z})_{MGTrel} = \sigma_{\bar{Z}}(\delta_p) \cong \sqrt{\sigma_Z^2(\Delta_p) - [\sigma_Z^2(\Delta) - \sigma_Z^2(\delta)]}, \quad (96)$$

where the correction to $\sigma_Z^2(\Delta_p)$ is the difference between the overall absolute and relative error variances for the composite scores.

MIRT

As previously discussed, the conditional composite score distribution, $\Pr(Z = z|\boldsymbol{\theta})$, can be obtained using Equation 60 under either the UIRT or the MIRT framework. The standard deviation of the conditional distribution is the CSEM for the composite, which is equal to the square root of Equation 62. For graphical representation of composite-score CSEMs, especially under MIRT, it is common to use the true composite score, $\zeta_{\boldsymbol{\theta}}$, as the conditioning variable. However, the relationship between $\boldsymbol{\theta}$ and $\zeta_{\boldsymbol{\theta}}$ in MIRT is not strictly one-to-one, leading to challenges in visualization. This is because of the infinite possible combinations of component θ's that yield the same composite true score. Thus, multiple CSEM estimates can be associated with each true composite score. To address this issue, various strategies can be employed, such as computing the arithmetic mean for each true composite score or applying a polynomial regression to obtain single-valued smoothed estimates.

A MIRT model typically provides better fit than a UIRT model when a test is composed of multiple subdimensions such as content areas. Oftentimes, however, a UIRT model is judged to be robust to a moderate violation of the unidimensionality assumption and more feasible to use in practice because of its computational simplicity and unambiguous interpretation of results. When applied to mixed-format tests, W. Lee et al. (2020) observed that the UIRT and MIRT models generally agreed closely in terms of the patterns of CSEMs; however, when aggregated to obtain the overall SEM and reliability, results tended to be quite different as the data became more multidimensional.

Summary and Other Issues

The procedures discussed in this section are summarized in Table 5.3, categorized based on their respective model, score types, assumptions about replications, and item format. Not covered in the preceding section are methods for estimating CSEMs for scale scores converted from composite scores. Composite scale-score CSEMs can be derived using the compound binomial model (Brennan & Lee, 1999), compound multinomial model (W. Lee, 2007), or MIRT (W. Lee et al., 2020).

When raw scores or IRT proficiency estimates are the primary focus, estimating CSEMs is relatively straightforward, and the general shapes of these CSEMs are highly predictable. However, in many cases, raw scores or IRT proficiency estimates are transformed into scale scores for reporting. In such cases, expressing CSEMs on the metric of a reported score scale provides more informative results. Previous research suggests that the overall pattern of scale-score CSEMs along the score scale tends to follow the changes in the slope (or first derivative) of a transformation function (Brennan & Lee, 1999; Feldt & Qualls, 1998; Kolen & Lee, 2011; W. Lee et al., 2000, 2020). A steeper slope at a raw-score point corresponds to a wider range of adjacent raw scores converting to a broader range of scale scores, leading to greater variability of scale scores over repeated measurements. Conversely, a flatter slope is associated with a wide range of raw scores converting to a smaller range of scale scores, resulting in less variability of scale scores over replications. It is advisable in practice that the relative magnitude of scale-score CSEMs be examined, especially near the score points that are critical for decision-making, such as cut scores for licensure and certification exams.

The delta method approximation can serve as a general procedure for estimating scale-score CSEMs because it can be used jointly with almost any estimates of raw-score CSEMs. In most applications in educational testing contexts, the delta method requires an approximation of the first derivative function through a *fitted* conversion table to obtain a continuous, differentiable function. A high-degree polynomial model is often chosen due to its flexibility and ease of differentiation. However, the performance of the delta method with polynomial approximation may be contingent on the characteristics of a conversion table. For example, if a conversion table is characterized as many-to-one, the polynomial approximation of slope may

Table 5.3 Estimators of Conditional Standard Errors of Measurement Based on a Single Test Administration

Procedure/Model	Score Type	Item Type	Assumption
Thorndike	Raw score	Both	Essential tau-equivalent split halves
Mollenkopf	Raw score	Both	Essential tau-equivalent split halves
Binomial error model	Raw score Scale score	DI	Randomly parallel
Compound binomial error model	Composite score Composite scale score	DI	Stratified randomly parallel
Multinomial error model	Raw score Scale score	PO	Randomly parallel
Compound multinomial error model	Composite score Composite scale score	PO	Stratified randomly parallel
Univariate GT	Raw score: Absolute error Relative error	Both	Randomly parallel
Multivariate GT	Composite score: Absolute error Relative error	Both	Stratified randomly parallel
Strong true score model	Raw score Scale score	DI	Randomly parallel
Unidimensional IRT	Raw score Scale score ML estimator	Both	Strictly parallel
Unidimensional IRT	EAP estimator	Both	Undefined
Multidimensional IRT	Composite score Composite scale score	Both	Stratified strictly parallel
Delta-polynomial method	Scale score Composite scale score	Both	

Note. DI = dichotomous item type; PO = polytomous item type.

be questionable. In that case, estimation procedures that leverage the conditional scale-score distribution may be deemed preferable, although further research is needed to substantiate this preference.

Virtually all estimation procedures for scale-score CSEMs discussed in this section assume a *constant* raw-to-scale score conversion table across hypothetically replicated parallel forms and for every test taker in the population. In real-world testing situations where multiple alternate forms of a test are developed and used, the concern is centered around the interchangeability of scores across these forms, which may vary in difficulty. Equating plays an important role in this context, and an outcome of such equating is a raw-to-scale score conversion table *specific* to each test form. The resulting scale scores are considered interchangeable across different forms if equating is trustworthy.

The need for equating to generate different conversion tables for each hypothetical replication in estimating scale-score CSEMs has not yet been systematically investigated in the literature. Obviously, the assumption of strictly parallel forms for the IRT procedures does not require equating because forms are effectively fixed, and any score differences are solely attributable to random error. In contrast, the assumption of randomly parallel forms in binomial and multinomial procedures, at least in the context of estimating CSEMs, does not require every test taker to take the same set of items on each replication. This implies that a test form, as a collection of specific items, is not clearly defined under this conception. If an additional restriction is imposed to allow for the same random sample of items to be used for all test takers in each replication, the resulting forms will have equivalent difficulty levels because all items selected for each replication are of equal difficulty for any given test taker. Therefore, under the randomly parallel forms assumption, equating is deemed either impossible or unnecessary. Further exploration of this issue may be warranted.

While the overall and conditional SEMs by themselves provide useful information about the amount of error in test scores, they are frequently used in conjunction with intervals. The *Standards* (AERA et al., 2014), especially Standard 6.10, recommend that score precision be depicted by error bands using the SEM. There is a voluminous literature on various types of interval estimation procedures in relation to different types of scores. Space does not permit any systematic review of the topic, but the following two principles merit consideration. First, the CSEM is usually favored over the overall SEM when constructing intervals for each individual test taker. Second, the endpoints of an interval should preferably be on the score metric used for reporting (e.g., scale scores). For example, W. Lee et al. (2006) recommended using CSEMs rather than the overall SEM based on a simulation study. They also proposed an endpoints conversion method for constructing intervals for nonlinearly transformed scale scores from raw scores, demonstrating superior performance compared to a normal approximation method.

As a final note, if a test has a multidimensional structure by design, such as a table of content specifications or mixed item formats, it is strongly recommended that the CSEMs for the total scores be estimated using one of the composite score models. Ignoring the fact that domains are stratified and fitting noncomposite models to the total scores could introduce bias in the estimated CSEMs, with larger bias as data become more multidimensional (W. Lee et al., 2000).

RELIABILITY OF CLASSIFICATION CATEGORY SCORES

When using a test score or composite score for categorical classification decisions (e.g., pass/fail), the *Standards* (AERA et al., 2014) recommend estimating the consistency of classifications across two replications of the same measurement procedure.

Classification errors occur when decisions are based on test takers' observed scores, which contain measurement error. In this context, the interpretation of test scores primarily considers each test taker's observed score relative to a standard or cut score, rather than the performance of other test takers. Since Glaser (1963) introduced the notion of criterion-referenced interpretation of test scores, extensive research has been devoted to developments of indices for measuring the precision of classifications. Berk (1980) provided a comprehensive summary of the measures developed until 1980, and since then, new methods have emerged as a result of increased use of IRT and complex assessment types.

This section reviews procedures based on threshold-error loss for assessing consistency of classifications. These methods primarily focus on the extent of consistent classifications with respect to the cut score, treating all false classifications as equally serious. By contrast, squared-error loss methods, such as those proposed by Livingston (1972), Brennan and Kane (1977a, 1977b), and Kane and Brennan (1980), consider misclassifying test takers with scores far from the cut score as more serious than misclassifying those with scores near the cut score. While both approaches offer meaningful insights into classification consistency, they address different aspects of the classification problem. Another group of procedures, not reviewed in this section, includes methods involving dividing the test into two parallel halves and stepping up the result to what might be expected for the full-length test (e.g., Breyer & Lewis, 1994; Woodruff & Sawyer, 1989).

The precision, or lack thereof, of classifications is often described in terms of classification consistency and accuracy in the literature. *Classification consistency* measures the degree to which test takers are classified in the same performance category based on two independent replications of the same (or similar) measurement procedure. A double administration procedure (Hambleton & Novick, 1973) using two parallel forms involves tallying the proportion of test takers assigned to each classification category on both administrations. By contrast, single-administration procedures rely on two expected (i.e., model-predicted) observed score distributions from two hypothetical replications of the test. The concept of replications, whether actual or hypothetical, leads to the perception of classification consistency as the reliability of classifications (W. Lee et al., 2002).

Classification accuracy assesses the agreement between classifications based on test takers' observed scores and classifications based on test takers' true scores (W. Lee et al., 2002; Livingston & Lewis, 1995). Unlike classification consistency, which compares two observed classifications, classification accuracy relates observed classifications to true classifications. In this sense, classification accuracy is sometimes referred to as decision validity (Berk, 1980; Hambleton, 1980).

Classification Consistency and Accuracy Indices

Suppose that test takers are classified into H mutually exclusive performance categories based on a set of $H-1$ cut scores on the raw-score metric, $c_1, c_2, \ldots, c_{H-1}$. For notational convenience, let c_0 and c_H denote the lowest and highest possible scores, respectively. The first performance category contains scores ranging from c_0 to $c_1 - 1$; the second category contains scores from c_1 to $c_2 - 1$; and so forth. Let $P_1, P_2 ..., P_H$ denote the observed performance categories in which each test taker is classified by comparison of their observed scores to cut scores. Further, let X_1 and X_2 be the random variables for observed scores on two administrations of the test that are independent and identically distributed.

The two most frequently discussed classification consistency indices are the agreement index ϕ and Cohen's (1960) kappa (κ) coefficient. The marginal agreement index ϕ is the percentage of test takers consistently classified in the same category on two independent replications:

$$\phi = \sum_{h=1}^{H} \Pr(X_1 \in P_h, X_2 \in P_h). \tag{97}$$

More specifically, the ϕ coefficient is the sum of diagonal elements of an $H \times H$ contingency table composed of joint probabilities or percentages of observed category classifications on two replications, denoted $\Pr(X_1 \in P_i,\ X_2 \in P_j)$. The κ coefficient adjusts for agreement occurring by chance and is given by

$$\kappa = \frac{\phi - \phi_c}{1 - \phi_c}, \tag{98}$$

where ϕ_c is the chance agreement: $\phi_c = \sum_{h=1}^{H} \Pr(X_1 \in P_h) \Pr(X_2 \in P_h) = \sum_{h=1}^{H} \left[\Pr(X_1 \in P_h)\right]^2$ because X_1 and X_2 are identically distributed.

Classification accuracy relies on a bivariate distribution of observed and true classifications. True classification involves determining a test taker's true categorical status based on their true score relative to cut scores expressed on the true score metric, $\lambda_1, \lambda_2, ..., \lambda_{H-1}$, with λ_0 and λ_H being the lowest and highest possible true scores, respectively. These true cut scores typically are set to be the same as the observed cut scores, but theory does not require doing so. Because test takers' true scores are unknown, individual true scores can be estimated, or an entire true score distribution for a population can be estimated or assumed. Let $\Gamma_h (h = 1,2,...,H)$ denote the true category of a test taker with true score τ such that $\lambda_{h-1} \le \tau < \lambda_h$. An $H \times H$ contingency table can be generated, containing the joint probabilities of observed and true classifications, $\Pr(\tau \in \Gamma_i,\ X \in P_j)$. This contingency table is asymmetric. The marginal classification accuracy index, denoted γ, is defined as the sum of the diagonal elements in the contingency table:

$$\gamma = \sum_{h=1}^{H} \Pr(\tau \in \Gamma_h,\ X \in P_h). \tag{99}$$

If the rows represent true categories and columns represent observed categories, the sum of the upper diagonal elements indicates the percentage of test takers classified in observed categories higher than their true category, known as the marginal false-positive error rate, symbolically expressed as

$$\gamma^{+} = \sum_{i=1}^{H-1} \sum_{j=i+1}^{H} \Pr(\tau \in \Gamma_i,\ X \in P_j). \tag{100}$$

Conversely, the marginal false-negative error rate represents the percentage of test takers whose observed categories are lower than their true categories and is defined as the sum of the lower diagonal elements of the contingency table:

$$\gamma^{-} = \sum_{i=2}^{H} \sum_{j=1}^{i-1} \Pr(\tau \in \Gamma_i,\ X \in P_j). \tag{101}$$

In essence, all estimation procedures discussed next involve, one way or another, constructing $H \times H$ contingency tables for consistency and accuracy. Nevertheless, they differ with respect to underlying assumptions ascribed to various models and applicable item types (dichotomous vs. polytomous) and score types (raw, composite, or theta estimates).

For most estimation methods, it is more convenient and informative to begin with estimating classification indices for each individual test taker or conditional on each level of true score. These conditional indices are then aggregated or averaged to derive the marginal indices. The subsequent sections primarily concentrate on the ϕ and γ coefficients to simplify matters.

Strong True Score Models

The first procedure is grounded in strong true score models. Huynh (1976) introduced an estimation procedure based on the beta-binomial model, the simplest form of a family of strong true score models. Later, Hanson and Brennan (1990) extended Huynh's approach to more general strong true score models. The conditional raw-score distribution, $f(x|\pi)$, can be modeled using either a binomial or a two-term approximation to the compound binomial model. The distribution of π for a population is either a two-parameter beta or a four-parameter beta distribution.

Let $f(x_1|\pi)$ and $f(x_2|\pi)$ represent the identically distributed conditional raw-score distributions on two replications of a measurement procedure. The conditional probability of being classified in the hth observed category P_h is

$$\Pr(X_1 \in P_h|\pi) = \Pr(X_2 \in P_h|\pi) = \sum_{x_1=c_{h-1}}^{c_h-1} f(x_1|\pi). \tag{102}$$

Due to the local independence assumption, the conditional joint distribution of X_1 and X_2 is

$$f(x_1, x_2|\pi) = f(x_1|\pi) f(x_2|\pi). \tag{103}$$

Given a set of $H-1$ cut scores, a *conditional* $H \times H$ contingency table can be generated based on the conditional joint probabilities. The probability of falling in the same *h*th category on two replications (i.e., diagonal elements) can be computed as

$$\Pr(X_1 \in P_h,\, X_2 \in P_h \,|\pi) = \sum_{x_1=c_{h-1}}^{c_h-1} f(x_1|\pi) \sum_{x_2=c_{h-1}}^{c_h-1} f(x_2 \mid \pi) = \left[\sum_{x_1=c_{h-1}}^{c_h-1} f(x_1|\pi)\right]^2 . \quad (104)$$

It follows that the conditional probability of consistent classifications is

$$\phi(\pi) = \sum_{h=1}^{H} \Pr(X_1 \in P_h,\, X_2 \in P_h \,|\pi), \quad (105)$$

which is referred to as the conditional agreement index. The marginal agreement index is obtained by integrating $\phi(\pi)$ over the entire distribution of π:

$$\phi = \int_{\pi} \phi\,(\pi) f\,(\pi)\, d\pi, \quad (106)$$

where integration is performed numerically using a set of quadrature points from an estimated true score distribution (e.g., beta).

Now, let us suppose that a true score (or a quadrature point) belongs to the *h*th performance level, $\pi \in \Gamma_h$. An accurate classification occurs when, based on observed scores, a test taker is classified in a performance category that is the same as the test taker's true performance category. Therefore, the conditional accuracy index is simply

$$\gamma(\pi) = \Pr(X \in P_h \,|\pi \in \Gamma_h), \quad (107)$$

which can be computed using Equation 102. The marginal accuracy index is then given by

$$\gamma = \int_{\pi} \gamma(\pi) f\,(\pi)\, d\pi. \quad (108)$$

The conditional and marginal false-positive and -negative error rates can be derived in a similar manner.

Normal Approximation Procedure

The normal approximation procedure (Peng & Subkoviak, 1980) was initially proposed as a simplification of Huynh's (1976) beta-binomial procedure. Research suggests, however, that this relatively simple procedure performs adequately across various situations and can accommodate different types of scores (e.g., S. Y. Kim & Lee, 2019, 2020; Wan et al., 2007). The crucial assumption of this procedure is that the observed scores from two replications follow a bivariate normal distribution with a correlation equal to test reliability.

Let $z(c_h)$ denote a z-score corresponding to each cut score c_h ($h = 1,2,\ldots,H$). This standardization enables us to use the convenient properties of z scores. The marginal percentages of test takers who are classified in the same *h*th performance category on two replications, $\Pr(X_1 \in P_h, X_2 \in P_h)$, can be determined using a cumulative standard bivariate normal distribution. The marginal agreement index ϕ is a sum of $\Pr(X_1 \in P_h, X_2 \in P_h)$ over all performance categories, as follows:

$$\begin{aligned}\phi &= \sum_{h=1}^{H} \Pr(X_1 \in P_h, X_2 \in P_h) \\ &= \sum_{h=1}^{H} \int_{z(c_{h-1})}^{z(c_h)} \int_{z(c_{h-1})}^{z(c_h)} \frac{1}{2\pi\sqrt{1-R(X)^2}} \exp\left[-\frac{x_1^2 - 2R(X)x_1x_2 + x_2^2}{2[1-R(X)^2]}\right] dx_1 dx_2, \qquad (109)\end{aligned}$$

where $R(X)$ is an estimate of reliability. Based on the fact that the marginal distribution of X_1 (or X_2) is univariate normal, the marginal category probabilities of $\Pr(X_1 \in P_h)$ and $\Pr(X_2 \in P_h)$ can be computed using a cumulative univariate standard normal distribution. The chance agreement and kappa coefficient can be computed accordingly.

Classification accuracy was not considered explicitly by Peng and Subkoviak (1980). Later researchers adopted the normality-based framework for estimating classification accuracy indices (e.g., S. Y. Kim & Lee, 2019). The assumption is that the marginal joint distribution of the observed and true scores follows a bivariate normal distribution. Since both observed and true classifications are involved, the true cut scores as well as the observed cut scores need to be standardized. For summed raw scores, $\mu(X) = \mu(T)$ and $\sigma(T) = \sigma(X)\sqrt{R(X)}$ in CTT. Thus, true cut scores are standardized as $z(\lambda_h) = [\lambda_h - \mu(X)] / \sigma(X)\sqrt{R(X)}$. The correlation of this bivariate distribution is $\sqrt{R(X)}$ because $\rho(X, T) = \sqrt{R(X)}$. The marginal accuracy index is given by

$$\begin{aligned}\gamma &= \sum_{h=1}^{H} \Pr(X \in P_h \mid \tau \in \Gamma_h) \\ &= \sum_{h=1}^{H} \int_{z(c_{h-1})}^{z(c_h)} \int_{z(\lambda_{h-1})}^{z(\lambda_h)} \frac{1}{2\pi\sqrt{1 - R(X)}} \exp\left[-\frac{\tau^2 - 2\sqrt{R(X)}\tau x + x^2}{2[1 - R(X)]}\right] d\tau dx. \qquad (110)\end{aligned}$$

A cumulative univariate standard normal distribution for each of the true and observed scores can be used to obtain the marginal category probabilities of $\Pr(\tau \in \Gamma_h)$ and $\Pr(X \in P_h)$, which then are used to fill the off-diagonal cells of the marginal $H \times H$ contingency table, $\Pr(\tau \in \Gamma_i, X \in P_j)$, for computing false-positive and -negative error rates.

The assumption of bivariate normality for the normal approximation method may not fully hold in reality. However, research suggests that this approach is quite robust to violations of the normality assumption (e.g., S. Y. Kim & Lee, 2020; Wan et al., 2007). Regarding the choice of reliability coefficient in Equations 109 and 110, it is generally preferred to use reliability coefficients that involve absolute error variance because the focus in classification contexts is on the discrepancy between test takers' scores and cut scores rather than their ranking. Peng and Subkoviak (1980) used KR21 with the normal approximation method, and Huynh (1976) also employed KR21 in the beta-binomial model framework, following Lord and Novick (1968), who demonstrated that reliability for the beta-binomial model is indeed KR21. Although the choice of reliability coefficient may have minimal practical impact, it is prudent to select one that is defensible for a particular application, especially paying close attention to the score type of interest.

Livingston and Lewis Procedure

The strong true score models (Hanson & Brennan, 1990; Lord, 1965) assume that n items are dichotomously scored and equally weighted. The Livingston and Lewis (1995) procedure is intended to extend the strong true score models to other complex types of test scores such as scores from polytomous items, composite scores, and even scale scores. For this somewhat ad hoc extension, they introduced the concept of *effective test length*, denoted here as $\tilde{n}$, which represents the pseudo number of equally difficult dichotomous items required to achieve the same reliability as the actual reported scores. (Note that the term effective test length carries a different interpretation compared to its usage in Equation 9). Letting X denote the actual test scores, the effective test length is computed as

$$\tilde{n} = \frac{\left[\hat{\mu}(X) - min(X)\right]\left[max(X) - \hat{\mu}(X)\right] - R(X)\,\hat{\sigma}^2(X)}{\hat{\sigma}^2(X)[1 - R(X)]}, \tag{111}$$

where $R(X)$ is a reliability coefficient for the original score type, and $min(X)$ and $max(X)$ indicate the minimum and maximum possible scores, respectively. Since $\tilde{n}$ can be a noninteger value, it is often rounded to the nearest integer. One of the strong true score models can be applied to this pseudo dichotomous data to compute classification indices as outlined in the section "Strong True Score Models."

The flexibility of the Livingston and Lewis procedure makes it one of the most frequently cited and used procedures in practice. Many comparison studies have found that results for the Livingston and Lewis procedure are comparable to those for other methodologies (e.g., S. Y. Kim & Lee, 2020; W. Lee et al., 2009; Wan et al., 2007).

Binomial and Multinomial Models

Subkoviak (1976) proposed a procedure for estimating classification consistency for a group of test takers under the assumption of dichotomously scored items. Unlike Huynh (1976), who utilizes the full feature of the beta-binomial model, Subkoviak's procedure does not entail estimating the true score distribution, although it assumes a binomial distribution for the conditional observed score distribution, $f(x|\pi)$. Instead of estimating the entire true score distribution, Subkoviak's procedure uses an estimated true score for each individual test taker. Subkoviak (1976) suggests using either a test taker's observed proportion-correct score or Kelley's (1947) regressed score estimate. The procedure computes the conditional agreement index for each person and then obtains the marginal agreement index by averaging the conditional results over all persons in the data set.

W. Lee (2005) and W. Lee et al. (2009) extended Subkoviak's procedure by: (a) proposing a procedure based on the multinomial error model (W. Lee, 2007) for polytomous items; (b) applying the compound multinomial model (W. Lee, 2007) to composite scores; and (c) considering classification accuracy, which was not addressed in Subkoviak (1976). The multinomial procedure, briefly introduced below, reduces to Subkoviak's model if items are dichotomously scored.

As previously defined in the section "Multinomial Error Model", for a person with $\boldsymbol{\eta} = \{\eta_1, \eta_2, ..., \eta_k\}$, the random variables, $Y_1, Y_2, ..., Y_k$, represent the numbers of items scored $a_1, a_2, ..., a_k$, respectively, which follow a multinomial distribution. The total raw score is $X = \sum_{i=1}^{k} a_i Y_i$. The conditional total score distribution given $\boldsymbol{\eta}$ is provided in Equation 80. The probability of being in the hth performance level is given by

$$\Pr(X \in P_h \mid \boldsymbol{\eta}) = \sum_{x=c_{h-1}}^{c_h - 1} f(x \mid \boldsymbol{\eta}). \tag{112}$$

For a person with $\hat{\eta}_i = y_i/n$, the conditional agreement index, $\phi(\hat{\boldsymbol{\eta}})_p$, is computed using Equations 104 and 105. These conditional agreement indices for all N persons are averaged to obtain the marginal agreement index:

$$\phi = \sum_{p=1}^{N} \phi(\hat{\boldsymbol{\eta}})_p. \tag{113}$$

Using observed score x as an estimate of true score, each person's true performance level is determined by comparing x to true cut scores. For a person with observed $\hat{\boldsymbol{\eta}}$ and $\lambda_{h-1} \le \hat{\tau} = x = \sum_{i=1}^{k} a_i\, y_i < \lambda_h$, the probability of accurate classifications is simply the probability of being classified into the hth performance category:

$$\gamma(\hat{\boldsymbol{\eta}})_p = \Pr(X \in P_h \mid \hat{\tau} \in \Gamma_h), \tag{114}$$

which is computed using $\hat{\boldsymbol{\eta}}$ in place of $\boldsymbol{\eta}$ in Equation 112. The gamma coefficient for a group of test takers is

$$\gamma = \sum_{p=1}^{N} \gamma(\hat{\boldsymbol{\eta}})_p. \tag{115}$$

The methodology discussed above can be extended to cases where the score type of interest is a composite score. The key difference is to model errors according to the compound binomial model, as discussed in detail by W. Lee (2005, 2007) and W. Lee et al. (2009). Brennan and Lee (2008) considered correcting for bias in classification consistency indices by proposing an estimator of true score that has the same variance as true scores.

IRT Procedures

Several procedures exist for computing classification indices under the IRT framework. For example, Huynh (1990) considered the Rasch model for classification consistency, with subsequent extensions to accommodate other UIRT models (W. Lee, 2010; W. Lee et al., 2002; Schulz et al., 1999; T. Wang et al., 2000). S. Y. Kim and Lee (2019) and W. Lee et al. (2020) further expanded the UIRT procedures to handle composite scores using assumptions of SS-MIRT and BF-MIRT models. While these procedures were primarily designed for tests based on summed scoring, those developed by Rudner (2001, 2005) and Guo (2006) are geared toward ML proficiency estimates. We first review Rudner's proficiency-based procedure, followed by methods tailored for summed scoring.

ML Proficiency Estimates

Rudner's procedure operates under the assumption that, for person *p* with an ML estimate $\hat{\theta}_p$, errors are asymptotically normal with a mean of $\hat{\theta}_p$ and standard deviation of $\hat{\sigma}_p = \sqrt{1 / \hat{I}(\theta,\hat{\theta})}$, where $\hat{I}(\theta, \hat{\theta})$ is a sample estimate of the test information function evaluated at $\hat{\theta}_p$. The expected probability of falling in the *h*th category is expressed as

$$\Pr(\hat{\theta} \in P_h \mid \hat{\theta}_p) = \Phi(\hat{\theta}_p, \hat{\sigma}_p, c_h) - \Phi(\hat{\theta}_p, \hat{\sigma}_p, c_{h-1}), \tag{116}$$

where $\Phi(\hat{\theta}_p, \hat{\sigma}_p, cut)$ is the cumulative probability below the theta cut score, *cut*, based on the normal distribution, $N(\hat{\theta}_p, \hat{\sigma}_p)$. Assuming errors are identically distributed for two replications, the probability of consistent classifications for each person is calculated as

$$\phi(\hat{\theta})_p = \sum_{h=1}^{H} \Pr(\hat{\theta}_1 \in P_h, \hat{\theta}_2 \in P_h \mid \hat{\theta}_p) = \sum_{h=1}^{H} \Big[\Pr(\hat{\theta} \in P_h \mid \hat{\theta}_p) \Big]^2 . \tag{117}$$

For a test taker with $\lambda_{h-1} \leq \hat{\theta}_p < \lambda_h$, the probability of accurate classifications is equal to the probability of being placed in the *h*th category:

$$\gamma(\hat{\theta})_p = \Pr(\hat{\theta} \in P_h \mid \hat{\theta}_p \in \Gamma_h). \tag{118}$$

The marginal indices of ϕ and γ for a group of persons are computed by averaging conditional indices as in Equations 113 and 115.

Guo (2006) proposed a modification of Rudner's procedure that relaxes the normality assumption. Instead of using the ML point estimate of θ and its standard error, Guo's approach utilizes the likelihood function of a test taker's item responses given item parameters to compute probabilities of classifications. Research has shown that both Rudner's and Guo's procedures tend to yield similar results (Guo, 2006; Wyse & Hao, 2012).

Summed Scoring

The formulation of IRT summed-score procedures is virtually the same as, or similar to, that of other procedures, particularly the strong true score models. However, there are two main distinctions: (a) the models used to derive the conditional observed score distribution (i.e., errors) and (b) the conditioning variables. For example, conditional distributions are defined for a UIRT model in Equation 50 or a MIRT model for composite scores in Equation 60. Subsequently, the steps outlined in Equations 102 through 108 can be applied to compute classification indices by replacing π with either θ or $\boldsymbol{\theta}$ depending on the model used.

Aggregation Methods

Estimating marginal classification indices requires a distribution of θ (or $\boldsymbol{\theta}$) for a population of test takers. While a discrete quadrature distribution is typically employed for a UIRT model, computational complexity grows exponentially with a MIRT model as the number of dimensions increases. To address this challenge, W. Lee et al. (2020) suggested alternative approaches for aggregating conditional indices. These include using: (a) quadrature distributions (D method), (b) individual proficiency estimates (P method), and (c) Monte Carlo simulation (M method). These three marginalization methods can also be applied to estimating CSEMs and reliability coefficients (see W. Lee et al., 2020).

The D method involves specifying a discrete set of combinations of quadrature points and associated weights for $\boldsymbol{\theta}$, often using a multivariate standard normal distribution. For example, with four dimensions and 21 quadrature points per dimension, there are $21^4 = 194{,}481$ possible combinations of quadrature points. Conditional agreement indices, $\phi(\boldsymbol{\theta})$, are computed for each combination and aggregated over the entire $\boldsymbol{\theta}$ distribution to obtain the marginal index: $\phi = \sum_q \phi(\boldsymbol{\theta}) w_q$, where the sum is taken over all combinations of theta values and w_q is the weight associated with quadrature combination q.

In contrast, the P method utilizes proficiency estimates as the conditioning variable for each person. Conditional indices are computed for each person with $\hat{\boldsymbol{\theta}}$, and a marginal index is obtained by averaging the conditional indices over the number of persons in the group: $\phi = \sum \phi(\hat{\boldsymbol{\theta}})_p / N$. The P method is typically computationally less intensive than the D method; however, results may be subject to the chosen proficiency estimator. The P method would be most useful when the purpose is to quantify classification errors for each individual test taker.

The M method combines elements of the D and P methods. Similar to the D method, it begins with an assumed distribution of $\boldsymbol{\theta}$, from which a large number of random deviates are drawn. For example, $N_r = 100{,}000$ combinations of theta values could be drawn from a multivariate standard normal distribution. Conditional results are estimated for each of the $N_r = 100{,}000$ simulated test takers. Then, similar to the P method, marginal results are computed by averaging the conditional results over the N_r test takers: $\phi = \sum \phi(\boldsymbol{\theta})/N_r$.

While these methods typically produce similar results (W. Lee, 2010; W. Lee et al., 2020), the M method is often preferred for MIRT models. It offers computational efficiency, eliminates unrealistic theta combinations, and yields conditional estimates—such as classification errors and SEMs—that are visually more interpretable when plotted.

Other Issues

The foregoing discussion on classification errors is restricted to cut scores established on untransformed score metrics. However, in many large-scale educational testing programs, cut scores are often set on reported score scales. Once a score scale is established for the initial test form, it is maintained across subsequent forms through test equating (Kolen & Brennan, 2014). In this scenario, there are at least two ways to estimate classification indices for scores on each test form. The easiest method would be to use the operational raw-to-scale score conversion table, in which the raw-score points that correspond to the cut scores on the scale-score metric could be identified. Alternatively, the conditional distribution of scale scores can be computed, as discussed in the section "Estimators of CSEMs." Then, the processes of constructing category probabilities and computing conditional and marginal indices, outlined in the section "Reliability of Classification Category Scores," can be applied.

The marginal classification indices provide useful information about the overall precision of classifying a group of test takers. However, they do not provide information about the precision of classifications at different score or proficiency levels. For this purpose, conditional indices along the score scale can be examined. Studies have consistently shown that classification errors tend to be larger near cut scores, as one might expect. Examples of exploring conditional classification indices are found in W. Lee (2010), W. Lee et al. (2002, 2020), and Wyse and Hao (2012). Wainer et al. (2005) considered a Bayesian method, involving the Markov Chain Monte Carlo procedure to produce samples from the posterior distribution of a test taker's proficiency estimate. The probability of passing, what they call the posterior probability of passing curve, is determined by counting the number of sampled proficiency values that exceed the cut score.

Most research to date on classification consistency and accuracy has focused on measurement errors that are attributable to items or forms. It seems natural to consider an extension of the current theories and methodologies to broader measurement situations. For example, classification decisions could be made based on test takers' responses to a set of essay prompts scored by raters. In such cases, an effective procedure should allow measurement errors attributable to essay prompts and raters to be modeled separately in constructing performance category probabilities. GT might prove to be a useful tool for developing such estimation models in classification contexts.

OTHER MODELS, AGGREGATION, AND PRECISION ISSUES

There are many additional aspects of reliability this chapter can only touch on.

Other Models and Assessment Types

Diagnostic Classification Models

With diagnostic classification models (DCMs), the latent traits are assumed to be categorical. This implies, typically, that there is a more limited range of possible latent values (e.g., master and nonmaster) than actual score values (raw scores, IRT proficiency estimates, or scale scores). In DCMs, the error variance and the latent trait variance are not independent, causing some complications in estimating reliability.

Using a replication definition of reliability, simulations are used to administer an assessment (the same form, a parallel form, or M. S. Johnson and Sinharay's (2018) definition of multiple assessments with an identical Q matrix and item parameters) repeatedly. What category test takers were assigned to each time was observed. Several authors have investigated different methods of estimating precision in DCMs, including M. S. Johnson and Sinharay (2018), who noted that simply reporting the estimated probability that an individual test taker is correctly classified may be misleading without considering the distribution of the categorical trait in the population. For instance, if 95% of a population are "masters," simply classifying every test taker as a master will result in 100% of the masters being correctly classified and 100% of the nonmasters being incorrectly classified.

Templin and Bradshaw (2013) introduced a measure of DCM reliability and compared it to an IRT measure. Replications for the DCM method can be obtained by repeated sampling from test takers' posterior distributions. For the IRT model estimate, the observed item responses provide an estimate of the latent trait, the standard error of which can be calculated using the test information function. Resampling via simulation can be used to obtain the distribution of the latent variable across the test-taker population, and the correlation computed between scores from two administrations can be used as an IRT reliability estimate. Comparing reliability estimates over the hypothetically repeated administrations, Templin and Bradshaw illustrated both theoretically and through the simulations that DCMs exhibited greater precision than IRT model-based estimates. Thompson et al. (2019) looked at DCM results for an operational assessment system reporting information, including reliability, at five levels and also found that the DCM reliabilities based on simulations approximating replications of testing were higher than those for CTT or IRT methods.

Structural Equation Models

Structural equation models allow for comparing various models for assessment data, which aids in the determination of an appropriate reliability coefficient in situations where one is uncertain about the dimensionality, tau equivalency, and other aspects of the data. Graham (2006) provided a hierarchy of models to estimate reliability, using structural equation models to test model fit to data, lamenting that

> many students and researchers in education and psychology are unaware of many of the assumptions required by a given statistical procedure, are unaware of how to test those assumptions, or are unaware of acceptable alternatives should those assumptions not be met. (p. 942)

Bacon et al. (1995) looked at the underestimation of reliability using coefficient alpha and omega across different conditions, advocating for weighted omega. Handcock and An (2018) provided an introduction to scale reliability within the context of confirmatory factor analysis and structural equation models, endorsing the use of McDonald's ω instead of Cronbach's alpha. S. B. Green and Yang (2008) derived a nonlinear structural equation model method to compute reliability for ordered categorical items. Raykov and Shrout (2002) presented a method to compute point and bootstrapped confidence interval estimates of reliability for weighted and unweighted composites with a general structure. Karimi (2015) illustrated structural equation models in estimating model-based reliability, including applications for the bifactor and mixed reflective-formative models, as well as covariate-dependent and covariate-free reliability.

Raykov and Penev (2010) outlined a latent variable analysis approach to provide both point and confidence interval estimates of reliability of group means, the overall reliability of group means, and conditional reliability for conditional and unconditional two-level models. Kano and Azuma (2003) discussed both the essential tau equivalence and the independence assumptions in a structural equation model environment for coefficient alpha, stating, "the independence assumption is more important than the essential τ equivalence assumption because dependency among unique factors can cause overestimation of the true reliability" (p. 147).

Oberski and Satorra (2013) addressed the "usual practice" of using error variances from an independent study as population variances in subsequent structural equation model analyses, showing how this may make the structural parameters' standard error too small. They provided an adjustment factor, illustrating its use with simulations and empirical data. Structural equation modeling continues to be a productive area of research related to precision coefficients.

Reliability Related to Longitudinal Data and Growth Models

Examining trends, such as an individual's scores over time or pretest and posttest scores to examine the effectiveness of a treatment, involves additional precision issues. Rogosa and Willett (1983) observed that when test takers have similar true growth changes, difference scores cannot adequately distinguish the amounts of true change among test takers. However, if the test takers differ in the amount of true growth, difference scores may indeed be adequate in distinguishing among test takers. Kane (1996) noted,

> The precision of change scores will depend on the intended interpretation of the change scores, in particular, on whether the interpretation focuses on the absolute magnitude of change for each individual or on the change for each individual compared with the average change in some reference group. (p. 368)

Thomas and Zumbo (2012) provided some support for the use of difference scores.

Raudenbush and Jean (2012) looked at reliability in the context of using value-added scores to examine teacher effectiveness, endorsing the use of confidence intervals over point estimates, and discussing multiple issues related to the use of value-added scores. Brennan et al. (2003) discussed various univariate and multivariate GT approaches to examine the reliability of group mean difference scores based on longitudinal data, from both norm-referenced and criterion-referenced perspectives, viewing the difference score as a composite score (the difference) of the two individual assessment scores.

Geldhof et al. (2014) examined reliability for data based on multistage sampling, concluding that level-specific calculations are appropriate when using multilevel data. Marcoulides (2019) examined reliability estimation in longitudinal studies, specifying the need to ensure that the approach used to model growth and measurement error aligned with the data being analyzed. Boyd et al. (2013) proposed an approach for estimating measurement error when students were administered three or more assessments, such as state assessments in consecutive grades. Their model allows for changes (increases or decreases) in knowledge and skills between administrations, as well as for the assessments to be neither parallel nor vertically scaled and to vary in their degree of measurement error.

Some research has examined change, difference, or growth scores in more complicated settings. Schuurman and Hamaker (2019) provided a preliminary model for some types of measurement error within the framework of autoregressive time series modeling, discussing how it could be relevant to both within-person and between-person precision. Tisak and Tisak (1996) used a latent curve approach to look at longitudinal models of reliability (see also Brandmaier et al., 2018).

DCMs have advanced to accommodate longitudinal data, examining how test takers' status changes over time. Madison (2019) defined and evaluated two reliability measures for longitudinal DCMs. The first measure is based on the idea of administering the same test repeatedly; the second is focused on the consistency of attribute mastery transitions, calculating the average most likely transition probability. Madison used both simulations and an empirical data set to study the two reliability metrics compared to a longitudinal IRT model. Consistent with the results of Templin and Bradshaw (2013), the longitudinal DCMs had higher reliability than the longitudinal IRT models and were easier to interpret.

Brennan et al. (2003) focused on longitudinal difference scores for matched students. In practice, schools likely have three groups of test takers when looking at growth across grades: students present (and tested) in both grades, students present in only the lower grade, and students present in only the higher grade. For the test takers having assessment scores in both grades, their scores tend to be positively correlated, which tends to decrease the standard error associated with the mean difference scores, whereas this is not the case when the test takers in the lower and higher grade samples do not overlap. In addition to larger error, conceptually using independent samples to look at differences across grades makes drawing conclusions about instructional effectiveness questionable (Brennan et al., 2003), a reminder that the data collection design needs to be considered in terms of the questions one wishes to consider.

These more complex designs and analyses tend to introduce additional sources of error that need to be considered in addressing precision issues, some of which may not be estimable. Haertel (2013) discussed many issues to consider related to teacher value-added scores, and Kane (2017) laid out issues related to using residualized student gain scores for value-added models, including bias in the individual student residual gain scores and the impact of this bias when averaging across students to obtain estimates of a teacher- or school-level effect. A discussion of possible "corrections" is provided, but with the caution that estimates reflecting "all sources of random error in the prior scores" (Kane, 2017, p. 10) are typically not available.

Adaptive Testing

Computerized adaptive tests (CATs) differ from other assessments in that the items a test taker is administered depend on the test taker's estimated proficiency. Multistage tests (MST) are typically of a fixed length, meaning a predetermined number of stages (and items) are administered to each test taker. Item or passage CATs generally stop after a set number of items are administered or after a fixed precision is reached. Although the standard errors of the proficiency estimates for each test taker can be used to obtain an approximate estimate of reliability, the standard errors will typically differ across test takers based on augmented stopping rules, such as stopping testing after a maximum test length is reached, regardless of the standard error of the current proficiency estimate.

A salient issue in adaptive testing is that basing which item or set of items a test taker sees next on how they responded to previous items suggests a violation of the item independence assumption, which IRT methods tend to rely on. The majority of methods used to estimate reliability for adaptive assessments use IRT methodology, though Divgi (1989) presented two methods to estimate reliability for a CAT without invoking IRT, and S. Kim and Livingston (2017) investigated a CTT-based procedure to estimate reliability for a MST, finding through their simulation that the method produced accurate results.

B. F. Green et al. (1984) stated, "Many psychometricians feel that devising a reliability coefficient for an adaptive test is inappropriate and misguided" (p. 352), but provided two alternatives, one being similar to an information function and the other using marginals. The authors suggested getting empirical reliability through test–retest designs. Beiser et al. (2016) computed test–retest reliability on an adaptive assessment of depression, concluding that the assessment provided "reliable screening." Haley et al. (2010) reported a study where test–retest reliability was computed using intraclass correlations across four CATs.

The *Standards* (AERA et al., 2014) suggest estimating reliability and precision information through simulations. Thissen (2000, p. 166) concurs, stating that

> because the entire system is used, including the item pool, the item selection algorithm, and the item exposure control system, such simulations may be expected to give accurate predictions about the performance of the CAT. . . . Simulation is the only situation in which the "real reliability" or "theoretical reliability" of a test can be determined.

Nicewander and Thomasson (1999) presented three test information–based reliability estimates for the Bayes modal estimate (i.e., MAP) and the ML estimate of proficiencies derived from direct definition, harmonic means, and Jensen's inequality, demonstrating that the latter two estimates were upper bounds for the true reliability. The results of multiple simulations found that ML and MAP provided nearly identical true reliabilities in all data sets and that all reliability estimates were within .02 of the true reliabilities. Segall (2001) introduced two methods designed to improve measurement precision of "a general test factor," one using a MIRT proficiency estimate and the other adaptively choosing items to maximize precision of a general proficiency score. Both methods were found to improve precision.

Nicewander (2018) transformed a reliability index to provide precision information for individuals, subgroups, and cut scores (conditional reliability coefficients), applying them to number-correct scores and theta estimates computed from number-correct scores, as well as theta estimates used as CAT scores. Seo and Jung (2018) compared three observed standard error marginalization methods (arithmetic mean, harmonic mean, and Jensen equality) in estimating empirical CAT reliabilities, finding that the Jensen equality method provided better accuracy and easy computation. Park et al. (2017) introduced a new method to obtain an analytic derivation of MST information.

Jodoin et al. (2006) studied test–retest and alternate forms reliability coefficients for several designs, including two- and three-stage MSTs, speculating that they would obtain higher coefficient values if there were more separations between their modules (i.e., if the medium module would have been less similar to the easy and difficult modules). Zhang et al. (2006) compared two IRT-based procedures for computing an aggregate reliability estimate using data from an MST assessment consisting of both MC and performance tasks. One procedure used an empirical ability distribution to estimate score variance, whereas the other assumed a normal distribution. The authors concluded that the two methods tended to produce similar results. The authors raised the issue that if the two methods provide quite different estimates, the question to consider is "whether reliability estimation in IRT scored tests, especially adaptive tests, should be sample-driven or sample-free" (Zhang et al., 2006, p. 13).

One issue to keep in mind in MST is that test takers are routed to a subset of items, so computing precision information over all items in a panel falsely inflates precision because each test taker is administered only the items included in a single path, not all possible items included over all paths.

C. Wang (2014) combined adaptive testing with hierarchical latent trait estimation, in hopes of increasing the reliability of hierarchical latent trait estimation. Two item selection methods were proposed, and both improved measurement precision over a unidimensional item selection method, particularly for short tests and when the correlation between dimensions was high.

Schmitt et al. (2010) investigated how speededness affected the reliability of proficiency estimation in a CAT, using simulation and two item pools, one "real" and one "ideal." The proficiency estimates became increasingly negatively biased as the CAT became more speeded in the real pool; these results were not replicated with the ideal pool.

Future precision research in adaptive testing will likely incorporate more complex IRT models, scoring with plausible values, and more complex assessment situations involving gaming and other simulations, where routing may be based on factors in addition to ability estimation.

Reliability Issues Related to Speeded Assessments

Operationally, some assessments are untimed, some are generously timed or partially speeded for at least some test takers, and still others have time limits imposed for practical reasons, such as when a test is administered in a school cafeteria and needs to be finished in time for lunch (for issues related to timing, see Margolis & Feinberg, 2020).

Estimating reliability for pure speed assessments from a single administration has been discredited for some time, with a consensus that administering two or more forms of an assessment to a group of test takers is the appropriate way to investigate precision (Anastasi & Drake, 1954; Cronbach & Warrington, 1951; Gulliksen, 1950), though attempts have been made to develop improved methods because of practical difficulties in obtaining data on two or more forms for a group of test takers (see Gulliksen, 1950). Cronbach and Warrington (1951) listed three conditions that result in inaccurate split-half or coefficient alpha reliabilities based on a single form when the assessment is speeded, including small variability in the number of items test takers have answered, when test takers have responded to all items they are likely to answer correctly within the time limit, and when the variation in the number of items responded to from form/occasion to form/occasion is small compared to the number of items responded to across test takers. According to these authors, if data from an assessment with a time limit do not meet one or more of these conditions, a split-half reliability estimate from a single administration might be usable.

In achievement and licensure/certification testing, relatively few assessments are pure speed tests, meaning there is not an expectation that lack of time is the sole, or even primary, reason for test takers not answering every item correctly. According to Cronbach and Warrington (1951), speededness is a factor to consider when a test taker's position in a group of test takers would change if the test taker were given additional time to respond to the assessment. Speededness violates the assumption of local independence of items, in that for those items a test taker does not have time to answer, the lack of time affects the test taker not answering the item correctly more than other characteristics, such as item difficulty, thus impacting the assumptions both GT and IRT methods make about local item independence for estimating coefficients.

Attali (2005) used simulations to illustrate the impact of speededness on reliability, assuming that test takers who were running out of time would engage in rapid guessing on their remaining items as opposed to leaving them unanswered, which tends to lower reliability because some items are answered correctly by chance, which is unlikely to be consistent across multiple test forms. Whether the speed at which a test taker works is consistent across alternate forms is another consideration.

The salient point is to what degree the test taker differences in scores are due to speed, or what proportion of total score variance is "speed" variance (Attali, 2005; Anastasi & Drake, 1954). When assessments are slightly speeded, or only a small portion of total variance is due to speed factors, internal consistency coefficients may not be too misleading. However, with the increase in computerized testing and the collecting of timing information, such as the amount of time a test taker spends responding to each item, more sophisticated methods of addressing speededness are possible, such as explicitly including timing information in estimating proficiency. Petscher et al. (2015) used a conditional item response model that incorporates response time as an item parameter in the context of a reading fluency assessment. Their model subsumes the idea of independence at the individual level, but incorporates a joint relation between speed and accuracy at the population level. The authors found their model improved the precision of student scores by an average of 5%. The increase was not uniform, however, having the most improvement for higher proficiency students. Other models, such as those introduced by van der Linden (2009) that incorporate latency, could also be used.

Aggregation Issues

Reliability of Group Means

The reliability coefficients dealt with in this chapter have concentrated on individual test takers and precision related to normative or criterion-referenced issues related to individuals. There are, however, instances where the unit of interest is at the aggregate level, for example, looking at a classroom. Examining trends, such as a cohorts' set of scores over time, or change scores, such as pretest and posttest scores to examine the effectiveness of a treatment, involves additional precision issues. Geldhof et al. (2014) examined level-specific (e.g., individual and group) reliability using three estimates and simulated and real multilevel data, finding support for the use of level-specific reliability estimates.

It is well known that the variability of, for example, individual fifth graders' assessment scores in a school district is generally greater than the variability of the average scores of the fifth-grade classrooms. Brennan (1995) used GT to illustrate, however, that the error variance for individuals could be less than the error variance for groups under certain conditions, depending on whether one considered persons and items random or fixed in terms of the universe of generalization. According to Brennan (1995), "Aggregation may well lead to a sizable decrease in error variance, but this can be very misleading if an investigator fails to take into account the corresponding decrease in true (or universe) score variance" (p. 395).

Reliability Generalization

Reliability is not a constant property of an assessment, but may vary based on context, such as a particular sample of test takers or number of raters, and the definition of a replication. Reliability generalization (RG) is basically a meta-analysis of reliability coefficient estimates across different studies, with the goal of estimating an average reliability across studies, as well as investigating the variability in the reported estimates and identifying factors that seem to be influencing the reliability estimates.

The importance of RG studies is that many researchers do not provide reliability estimates based on their particular context, but instead rely on, perhaps, a publisher's technical documentation for their reported reliability information, even though their context may differ in ways likely to affect precision. Although not as ideal as having precision information calculated on their particular data, RG analyses may provide some information on whether "borrowing" reliability information from one context and assuming it applies in another is likely to be reasonable (see Vacha-Haase et al., 2002, and Whittington, 1998).

There are different ways to implement an RG study and different factors to consider as possible influencers on estimated coefficient values. Holland (2015) examined 107 peer-reviewed RG studies, finding most analyzed coefficient alpha coefficients. He provided a set of guidelines for both conducting and reporting RG studies. Sánchez-Meca et al. (2019) also developed a checklist to assist researchers in reporting reliability generalization results. Henchy (2013) examined 64 RG studies, comparing the recommendations around such studies to how they were actually conducted, finding that some recommendations, such as conducting a priori power analyses, were not followed. Those conducting RG studies need to attend to issues such as whether to weight study results by test-taker sample size and how to treat a possible lack of independence of reliability estimates where a given study has reliability estimates from various subgroups, subscales, or multiple methodologies.

Precision Issues Related to Raters

Rater Reliability

When the scoring of a test taker's responses is impacted by who does the scoring, this source of variability should be incorporated when computing reliability. Rater impact is often a matter of degree and encompasses various situations, such as a teacher and a teacher's aide interpreting penmanship differently to various trained raters scoring constructed response items using an elaborate rubric on a statewide accountability assessment. The more likely different raters are to assign the same score to a test taker's response, and the more likely the same rater is to assign the same score to the same response, the less rater variability contributes to error variance. Good rater training, well-defined scoring rubrics, and monitoring to ensure raters are interpreting the rubrics and responses the same way help increase rater consistency.

GT is frequently used to model the variability of raters, as well as to estimate precision information over various features, such as the number of raters and the number of prompts or tasks administered. However, it is not always clear how well the results of these studies generalize to a particular context. Studies done by publishers during the initial administration of an assessment may include expert raters, training materials that are still under development, and prompts and rubrics that are still being tweaked, as well as administration conditions, such as time limits, that are still being determined, and they may or may not include test takers who are motivated. All of these factors can impact the GT precision information and make it less relevant for a later context. In addition, under operational conditions, raters are likely not all at the expert level, and raters may not be assigned randomly but rather by convenience, such as which raters happen to be available for a particular shift when a particular set of responses are available for grading. For local GT analyses, existing data are often searched for a small subset of data that align with a standard GT design, such as a fully crossed rater by test taker nested within prompt design, to gather precision information, including rater effects.

Rater scores are often compared through either a rank ordering of responses (correlation between rater scores over a group of test-taker responses) or a match of actual assigned scores (often contingency tables focusing on exact or adjacent matched scores, frequently corrected for chance agreement). Rater severity and leniency; the tendency of a rater to assign high or low scores, respectively; and proclivity of a rater to use the whole range of scores or not are perhaps the most frequently cited rater effects. The scoring process logistics, such as scoring all responses to a single prompt versus scoring all responses by a single test taker, can mitigate some rater issues, such as the halo effect.

Studies should account for variability within and across raters. Models incorporating rater effects include IRT models (Linacre, 1989; Lunz et al., 1990; Verhelst & Verstralen, 2001), a hierarchical rater model (Patz et al., 2002), a signal detection theory model (DeCarlo, 2010; DeCarlo et al., 2011), Yao's rater model (Z. Wang & Yao, 2013), and Longford's (1994) model-based approach for multiple raters of essays. Recent research includes that of Zapf et al. (2016), who investigated confidence intervals for nominal data and Fleiss's kappa and Krippendoff's α for interrater reliability in multiple scenarios, finding that bootstrapped confidence intervals provided coverage probability close to the theoretical ones. Choi and Wilson (2018) proposed a generalized linear latent and mixed model approach, combining IRT and GT, which allows ML estimation of individual random effects and variance components for generalizability coefficients. Gianinazzi et al. (2015) asked the same raters to rate the same responses at two points in time, looking at both intrarater and interrater reliability. Abdalla (2019) looked at the scoring of prompts reused over time under different conditions of rater drift, looking at the percentage of exact agreement and Cohen's kappa as interrater agreement measures, and the paired *t* test and Stuart's Q as marginal homogeneity measures, under two data frameworks: the generalized partial credit model and latent-class signal detection theory model.

Haertel (2006) discussed several rater issues, including what is the true score of test takers on a prompt where raters score responses. Is it the average score across replications of raters? Or across a test taker's repeated responses to the prompt, if the test taker could be repeatedly administered the same prompt? Or is it based on the rubric and not on rater's assigned score? Haertel (2006) also mentioned issues related to scoring with a finite rubric, such as 0 to 3, and the implications for test takers with true scores at the extremes. For example, a test taker with a true score of zero can only have measurement errors in one direction. Haertel (2006) stated, "Unless all raters assign the identical score, the mean observed score over raters is a biased estimator of the essay's true score. Thus, from this perspective, true score cannot be equal to the expected value of observed score" (p. 102). Haertel emphasized the need for precisely specifying the definition of true score and error, as well as the model being proposed, to minimize ambiguity and determine the appropriateness for a particular context.

Haertel (2006) also raised the issue of what score is assigned when multiple raters score the same response. Typically, multiple scores are averaged only if the scores are identical or fall within some predetermined range of each other; for example, adjacent scores may be averaged, but scores that are not identical or adjacent may be sent to resolution. Resolution may involve ignoring the previous rater scores and using only a score assigned by a more expert rater. Other rules may involve keeping the original rater score that most aligns with an expert score, discarding the more disparate original rater score. It is also the case that any averaging of ratings is likely to result in fractional scores that do not directly align with the scoring rubric. Both fractional scores and resolution scores impact variability due to raters. Haertel (2006) suggested that additional research on how to incorporate these cases into a rater model is warranted.

Variability across raters can be examined by assigning multiple raters to score the same responses. Variability within a rater is somewhat more difficult. A rater may remember seeing a particular response previously, so having a rater score a single response multiple times as independent scores may be challenging, as well as resource intensive. Examining how consistently a rater scores “essentially identical” responses might provide some insight, though the issue then becomes what is essentially identical and under what conditions (time of day, position in a stack of responses, before or after a higher or lower scoring response, length of time between scorings, etc.). Though raters are typically not the largest source of error in the precision of rater-scored responses, lagging behind task/prompt variability, rater variability is still a source of unreliability to minimize.

Automated Scoring Issues

Machine scoring, or artificial intelligence (AI) scoring, involves training a computer, through natural language processing or other methodology, to score essay/constructed responses by machine, rather than human raters. AI scoring is appealing because, in theory at least, it is more efficient and returns scores faster and more cost-effectively than human raters. Note, however, that current AI engines tend to require some training on each individual prompt, so the benefit in efficiency is largely manifest at scale when large numbers of test takers respond to the same prompt. (See, however, Attali, 2011 and Foltz et al., 2013 for a generic model; Foltz et al., 2013 reported the generic AI rater as having about 10% lower reliability than the prompt-specific rater). Although most of the research literature in AI seems to focus on essay scoring, there have been studies on other types of constructed response items (see Shermis, 2015, who looked at machine scoring short-answer items).

Whether assessment responses are scored solely by a single AI engine, multiple AI engines, or a combination of an AI engine and one or more human raters, precision coefficients are impacted. The most common way AI scores are currently evaluated is to look at the consistency between scores assigned by human raters to a group of test-taker responses correlated with the scores an AI engine assigns, where the AI engine scoring is typically trained to maximally predict the human scores based on features that may or may not resemble the human scoring rubrics. Various coefficients, including Pearson correlations between the two sets of scores, exact and adjacent agreement, and kappa, have been used (see, for example, Attali, 2013, who examined correlations between AI and human scoring across multiple studies, typically comparing human–human correlations to AI–human correlations). Attali (2015) examined different weights applied to AI features to find the optimal weights for reliability and validity, as opposed to the traditional use of predicting scores human raters would assign, reporting that the feature weights that predicted human rater scores were different from those that emphasized reliability coefficients.

Attali and Burstein (2006) discussed issues related to using human rater and AI agreement as reliability, including when data on only a single prompt are used. While a human rater may or may not assign the same score to a test-taker response presented twice, an AI engine would be expected to return the same score, assuming that no intervening tweaking or training of the engine occurred, suggesting it may be preferable to compare scores across multiple prompts rather than a single prompt.

Attali et al. (2013) looked at increasing the reliability of combined human and AI scores, both by decreasing the overlap in the features rated and by increasing the reliability of the human scores. Both methods proved effective. Attali (2011) reported that a test–retest reliability for a single human rater was .53, while that of a generic AI rater was .80. However, in the same way that all prompts, rubrics, administration conditions, samples, human rater training protocols, and so on vary, so do different AI engines, so while these results may be informative, they may not generalize to other contexts.

One question that arises in considering AI scoring is what actually constitutes a replication. Sending a specific response through the same AI scoring engine twice does not reflect the same variability as sending the same response through human scoring twice, with either the same or a different rater. The same test taker could be administered the same prompt or form twice and both responses could be sent through the AI engine, but this is not really a replication to investigate the AI engine because the within-person variability confounds the interpretation. Perhaps training the AI engine twice on two separate sets of training or calibration papers and then running a new set of responses through both engines and comparing the paired scores would be informative, but still would not really align with the typical human rater analyses.

Reliability of Behavioral Observations

Another setting where raters are involved is in observations of behavior, such as observers in a classroom setting looking for incidences of certain behaviors, the absence or presence of which on the part of teachers and/or students is recorded. Individual rater errors consist of recording an occurrence when it is not manifest and not recording an occurrence when one is manifest. Issues in this context involve rater consistency (one rater observes a behavior and one does not) and rater accuracy (either the behavior did or did not occur, so both raters observing the same incident are not correct), and representativeness—if one wishes to make inferences about behaviors based on the observed/rated sample, how representative is the observation period? For example, is an observed teacher demonstrating typical behavior, or is the fact of being observed leading a teacher to be more likely to demonstrate a desired behavior?

Gwet (2012) discussed issues around interrater reliability in medical settings, raising the issue of whether two raters whose ratings correlate highly are “interchangeable” and to what extent we can extrapolate from raters in a study rating particular subjects to a broader context, emphasizing the importance of gathering precision information using raters similar to those that will be used operationally. Gwet (2014) also discussed ordinal measurement issues, when the distinction of whether raters agree or disagree may be a matter of degree as opposed to absolute, offering the illustration,

> Two raters A and B who rate the same patient as "Certain Multiple Sclerosis" and "Probable Multiple Sclerosis" are not quite in total agreement. But are they in disagreement? Maybe to some extent only. That is, with ordinal scales, a disagreement is sometimes seen as a different degree of agreement. (p. 16)

This suggests that a rank ordering of test takers, rather than a contingency table, might be preferred to examine consistency.

The reader is referred to Haertel (2006) for additional information on issues related to precision with behavioral observations, particularly his treatment of Rogosa and Ghandour (1991) and their framework for considerations of score accuracy.

Error in Reliability Estimates

Reliability coefficients are used to provide an indication of the uncertainty in measurements (see Frank, 2002, for the use of reliability coefficients to correct effect sizes), but reliability estimates also contain error. In the examples provided in W. Lee et al. (2025), coefficient values across the two contrived forms can be used as a measure of consistency, but in practice, it is uncommon for more than one administration to occur, and simulations or resampling procedures are often used to estimate standard errors and confidence intervals. While this information can be informative, some sources of error, such as test-taker consistency across alternate forms or separate administrations, are not fully considered.

Several researchers have investigated issues around error in reliability and precision estimates. Andersson and Xin (2018) derived asymptotic variances for IRT marginal and test reliability coefficient estimates for both dichotomous and polytomous IRT models and demonstrated through simulations that the resulting confidence intervals had "good coverage" under several conditions studied, with the marginal reliability coefficient having somewhat lower sampling variability and larger bias than the test reliability coefficient.

Feldt (1990) provided methodology to create confidence intervals for the intraclass reliability coefficient, used when there is a single rater for individual responses, based on sampling theory. Ogasawara (2009) derived asymptotic distributions of several sample coefficients with and without stratification under nonnormality. Terry and Kelly (2012) focused on estimating the sample size needed to obtain narrow confidence intervals for composite coefficients. Bujang et al. (2018) looked at sample sizes relative to hypothesis tests for Cronbach's alpha coefficient. Zapf et al. (2016) looked at confidence intervals for interrater reliability estimates for nominal data.

When using an estimate of precision, particularly to inform a high-stakes decision, it is important to consider the precision of the estimate as well. This is likely to be of particular concern in GT where variance component estimates may be based on a sampling of very few levels from an infinite population (see Brennan, 2001c).

Practical Guidelines

What Reliability Computations Are Most Appropriate for What Context?

As has been discussed throughout this chapter, there are a variety of methodologies that can be used to compute precision information. According to the *Standards* (AERA et al., 2014), the reliability methodology selected for use in a particular context should "be consistent with the structure" of the assessment, providing the example that an assessment with considerable multidimensionality would indicate that a composite score reliability be used. It is also recommended that precision information be computed for the scores on which decisions are going to be made. For example, if equated and rounded scale scores are used for placing students into their first-year college math class, precision information provided on equated and rounded scale scores will be more appropriate than internal consistency information on raw scores.

E. Haertel (personal communication, March 8, 2019) advocated framing the issue of which reliability computations are most important by first considering what the intended use of the data is and what sources of error should be considered ("begin with a substantive definition of the intended universe of generalization"). This is then followed with selecting a data collection design that would populate the appropriate model. GT can often be extremely helpful here by explicitly identifying what factors one wishes to generalize over calculations (see Revelle & Condon, 2019). It is not always possible to collect the desired data, for example, administering the same assessment over and over to the same test takers. However, knowing what is desired can be helpful in choosing what available methodology to substitute.

To explore the various reliability statistics that can be used to represent consistency and inconsistency of scores, two large-scale examples are provided by W. Lee et al. (2025). For each of the examples, multiple reliability coefficients, CSEMs, and classification consistency and accuracy coefficients are provided. In addition, results are provided for number-correct raw score, IRT proficiency, and scale-score metrics for various statistics and indices. The purpose of providing these examples is to illustrate similarities and differences in values across multiple reliability statistics and to reinforce the importance of identifying details about how a reliability estimate was computed, instead of simply reporting "the reliability was [some value]."

W. Lee et al. (2025) demonstrated through examples that different reliability coefficients calculated on a single set of data yield different values. This is largely because the different coefficients make different assumptions. Simply running a multitude of analyses and choosing to report the highest values is counterproductive. Woodruff and Wu (2012), in a paper comparing coefficient alpha with multiple other coefficients, reminded us to remember what each reliability coefficient indicates and that we should be reporting the coefficient that best aligns with the inferences we plan to make.

How Do I Increase Reliability and How Much Do I Need?

The *Standards* (AERA et al., 2014) do not provide criteria for minimally acceptable reliability coefficients. Occasionally a request for proposal, a government organization, or an author (see Dimitrov et al., 1999) may specify a particular value of reliability that should or must be achieved, but there is rarely any rationale provided for the chosen value or any criteria for the conditions under which the value must be achieved (such as test-taker sample size, whether under operational or special study conditions, or the metric or type of coefficient). Reliability coefficients appearing in technical manuals may provide some indication of the magnitude thought acceptable, but typically these are lacking in detail about the data collection, and there is no evidence that these values are "adequate" other than that the test publisher thought they were sufficient to publish the assessment. Ellis (2013) commented, "The absence of such standards is a serious gap in existing behavioral research methods and leads one to ask why reliabilities are routinely computed if their acceptable values are unknown" (p. 16). Wainer and Thissen (1996) brought up the related issue of the practical impact of small differences in reliability coefficients.

Kane (1996) advocated judging how adequate a precision level is for one's intended use by considering the measurement error in relation to the "tolerance for error" in a specific situation:

> The level of error that is tolerable in a particular context is determined by the interpretations to be applied to the measurements and the uses to be made of the measurements. To the extent that errors do not interfere with intended interpretations or uses, they are not a serious problem. (p. 356)

This leads to reframing the question "How much reliability do I need?" to "How much error (uncertainty) can I tolerate in the decisions I need to make with these data?" Practitioners may not be concerned with relatively large CSEMs occurring far away from a cut score. However, when using an assessment as a component in a decision, such as a test score, a grade point average, or soft skill measures in determining college admissions, it may be much more difficult to articulate the amount of "tolerance for error" in a component.

Different factors impact different precision calculations differently, again depending on what the specific calculations account for. For example, differences in test specifications across test forms have no impact on computing a reliability coefficient based on only one administration of a single form, but would have an impact on a coefficient involving parallel forms or on analyses involving different forms over time. Alternate forms reliabilities will tend to increase as the forms are more similar.

Internal consistency measures increase as the items within an assessment have higher discrimination values. To rank order test takers according to some characteristic or trait, items that all test takers or no test takers answer correctly provide no information. Jodoin et al. (2006) reported that higher coefficient values for MST tests would be obtained if there was more separation between the easy, medium, and difficult modules.

The more objectively an assessment or item is scored, the more likely the reported scores are to be consistent and the less the scoring process is likely to contribute to error variance. The way an assessment is scored can affect reliability; Pugh and Brunza (1975) stated that the reliability of a test increased from .57 to .85 when comparing traditional number-correct scoring to a confidence scoring method. In addition, when multiple components are used to form a composite score, giving more weight to the more reliable components tends to increase reliability of the composite. Kane and Case (2004), for example, addressed how to weight two components of a composite score to maximize correlation with desired true score (defined as a specified weighted composite of component true scores) when the observed component scores have different reliabilities.

Another factor that can impact reliability coefficients is the metric used, as shown in tables associated with the two examples reported by W. Lee et al. (2025). There are multiple ways in which scale scores, theta scores, and cut scores could be set, which would also potentially influence precision calculations.

One factor that seems to be generally accepted is that longer assessments have more precision than shorter assessments because they provide a larger sample of tasks or items. However, there are caveats to keep in mind. One is that increases in precision due to increasing assessment length is likely most effective when the initial test is short. Adding 30 items to a 10-item test typically will increase reliability more than adding 30 items to a 200-item test, other things being equal. "Other things being equal" is an important caveat. Adding poor items (where "poor" may indicate inferior statistically, or in terms of alignment to a test blueprint, or items of such novel format that test takers are confused, or items that make the test so long fatigue and lack of motivation become issues) are more likely to lower precision than increase it. As Thorndike (1951) stated, "It has been shown that the reliability of many tests could actually be increased by omitting a number of items on the test" (p. 602).

In addition to increasing the number of items, the length of a multiple-choice assessment may also be considered in terms of increasing the number of options per item, again assuming the additional options are viable and of equivalent quality. The relationship of item sets to reliability may be somewhat more complex because of local dependence. Adding additional items per item set, where all items refer to the same set of stimulus material, is unlikely to have the same impact on reliability as adding additional item sets. According to Livingston (2018), "A test taker who has difficulties with that particular stimulus will have trouble with all the questions about it. To improve alternate-forms reliability, we need to increase the number of item sets" (p. 16).

Wainer and Thissen (1996) provided an example for which a test of shorter passage sets needed 43% additional items and a test of longer passages need 65% more items for each to obtain the reliability of a test of 40 stand-alone items. Lawrence (1995) also looked at the issues of local item dependence and the effect on reliability, finding that item sets effectively reduced test length: "The effect of item sets is revealed by a systematic discrepancy between the internal consistency estimates, which treats items based on the same passage as independent, and the parallel-form estimates" (p. 13). Passage sets may be an excellent way to assess certain content, but the introduction of item dependence affects some precision coefficients through a violation of assumptions.

When scores involve raters, adding additional independent raters tends to increase precision, again assuming the additional raters are of similar quality as the original raters. The question of whether adding additional prompts/items or additional raters would be most likely to improve reliability in a particular setting requires the total design to be considered. GT studies allow one to estimate the influence different factors would have on certain types of reliability coefficients, such as whether adding additional items or prompts, raters, clusters of students per school or more schools, and so on would provide a better return on investment in a particular context (see Sun et al., 1997). The design implemented for data collection impacts precision, especially from the standpoint of the intended universe of generalization. Whether a particular set of raters or prompts is considered fixed or random may have a substantial impact on the coefficient values. However, the goal is not to choose a design, such as assuming that raters are fixed, to raise a coefficient value, but to have the design reflect the context one is interested in.

Different data models, as well as different precision estimates, have different assumptions and incorporate different sources of variability into their definition of error. Therefore, selecting a coefficient because it is larger is not recommend. Coefficients should be selected to fit one's desired use and to align with the data collection design one implements.

Inappropriate administration conditions for an assessment can lower precision estimates by adding in additional error. These factors can range from inappropriate time limits, use of unfamiliar item formats or technology, inadequate motivation on the part of test takers to score their best, inadequate test directions, sloppy scoring, and a myriad of other conditions. Adequate instructions and practice exercises to ensure test takers understand what they are being asked to do tends to increase precision (see Thorndike, 1951).

For some consistency measures, the test-taker sample on which they are computed is a factor with large influence on the calculated value of the coefficient, such as the distribution of masters/nonmasters for a classification consistency measure or the overall distribution of test takers, such as how bunched or spread out the test takers are in terms of the underlying trait one is trying to estimate. The issue of restriction of range, when the possible value range is truncated through self-selection (for example, when only previously failing candidates are administered a retest or only high-proficiency candidates self-select to take an assessment) has been studied by various researchers. Sackett et al. (2002) examined the effect of different scenarios of range restriction on reliability coefficients, finding that the different scenarios resulted in different underestimates of criterion reliability. Fife et al. (2012) also looked at various selection ratios in relation to several reliability coefficients, including KR20 and test–retest, corrected and uncorrected, under varying selection ratios. These estimators were examined in terms of bias and precision, with test–retest reliability usually being the best estimator of reliability across conditions.

Improving precision has multiple perspectives. Deliberately testing a sample with an extreme range of ability may result in a larger calculated value of a coefficient for, say, technical documentation, but doing so influences only that particular administration, whereas adding additional items to the test will likely improve precision for every future administration. Similarly, narrowing the desired universe of generalization, from a GT perspective, by considering, for example, raters fixed may raise a value, but is inappropriate if those raters will not be scoring future responses. Knowing what factors may influence a particular reliability coefficient facilitates comparing values across assessments by considering the context under which each coefficient was calculated.

Effective Ways to Communicate Reliability Information

Once reliability information has been estimated, these values need to be communicated to those wanting to make decisions based on the scores. The statement "The reliability of Assessment A is .80" is virtually meaningless in terms of helping a user know how much confidence to place in the scores of interest. Without additional information, one cannot even be sure that a value of ".80" is better than a value of ".70" but not as good as a value of ".90," because one does not know whether the three values are for the same type of coefficient (or what sources of error were incorporated), or what sample of test takers was used in each instance, or other details, such as administration conditions, scoring, and so on. When presenting precision information, the relevant context must also be presented, so in addition to information about the actual coefficients computed, information on the assessment, the administration, the test-taker sample, the scoring, the reported score metric, and other pertinent details should be included. An internal consistency coefficient of a released practice form given under lax conditions with no stakes to students reported on number-correct scores and an alternate-forms coefficient under high-stakes standardized conditions to a self-selected sample reported on a scale score metric might be interpreted quite differently by someone planning to incorporate the assessment into a composite for decision-making, even if the coefficient values were similar.

Given there is sufficient detail to understand the context of the design and the coefficient calculations used, "uncertainty" needs to be communicated in a manner that facilitates understanding. A statement such as "A coefficient value of .75 suggests that 75% of the variance in scores is due to actual differences in ability while the remaining 25% is attributable to measurement error" is unlikely to be helpful to a lay audience. S. Johnson and Johnson (2009, p. 51) stated,

> If the public is to be educated about technical issues in assessment, and if reliability information is to be routinely published alongside examination results, then we need to decide which form of reliability measure would be the most appropriate one to use. There are basically two choices: a variance ratio (reliability coefficient), and a standard error of measurement.

According to C. Wang (2014, p. 454), "reliability has the advantage of being a compact measure of precision that has a fixed metric (bounded between 0 and 1) that is widely understood." Cronbach (2004) asserted that the SEM should take precedence as the most important piece of information to report, rather than a reliability coefficient. Patterson (1955), in an article dedicated to interpreting SEMs, discussed problems in making statements about SEM concerning true scores using the standard error of observed scores.

Walsh et al. (2014) illustrated reporting information graphically rather than as a single number. Raudenbush and Jean (2012) advocated the use of confidence intervals for value-added scores for teachers; they also discussed issues around interpretation, such as shorter intervals (e.g., 75% versus 95% confidence intervals) and the different interpretations one might make (such as "Where do I as a teacher stand among all teachers?" versus an administrator wanting to identify a lower group of teachers for training or a higher group for recognition).

Regardless of one's view on what measure of precision to report, illustrations and examples can help practitioners understand abstract concepts around uncertainty. For example, statements suggesting how many masters/nonmasters are likely to be misclassified, or how a test taker is likely to score if given the assessment a second time, or illustrating through simulation the distribution of obtained scores a test taker (simulee) of a given true proficiency achieves on two different CAT designs may be more impactful than simply reporting a coefficient value. Providing a comparison can also be useful to aid in interpretation, as Wood (2020) illustrated:

> When possible, documentation should include agreement statistics for two independent human raters and for a human rater with automated scoring. The public may be surprised if a scoring engine agrees with a human 65% of the time but may not be so surprised if two independent humans agree only 67% of the time on the same prompt. (p. 144)

FUTURE DIRECTIONS

In the years since the fourth edition of *Educational Measurement* was released, there has been progress in addressing reliability issues with more complex data models, with incorporating additional sources of error in coefficients, and with the development of new methodologies and coefficients. There has also been increased emphasis on the use of CSEMs over reliability coefficients, which this chapter hopefully reinforces.

As computing power and the integration of computer science and data analytics into measurement approaches continue to advance, more complicated simulation methods are available to examine reliability across more complex assessment models. As our models and assessments become more complex, so do our methods of estimating reliability. From fixed forms to adaptive tests to game-based assessment experiences, the concept of replication becomes more challenging. It may be relatively easy to imagine a test taker taking the same fixed form (or a randomly parallel or tau-equivalent form) a second time.

However, when it comes to adaptive testing and game-based assessment, with various possible paths involving the same or different pools, the scenario becomes more complex.

Simulations have become more necessary to estimate reliability, such as running the same simulees through alternate CAT pools multiple times. These simulations rest in part on having realistic models to generate test-taker responses, and we are not there yet. Too many simulations are conducted with models with too small of a random (or even some systematic) error component. We need new models of data generation along the lines of those used by Kolen and Harris (1987) and Davey et al. (1997) to more accurately reflect actual test-taker behavior if we are going to rely on the generated data for reliability information.

The previous editions of *Educational Measurement* have primarily focused on the reliability of summed scores from a single point in time (Haertel, 2006). This chapter extends its focus to include additional metrics such as IRT proficiency estimates and scale scores, recognizing the importance of reliability information aligned with the specific metric used for decision-making. In the future, additional metrics will likely become more prevalent in operational use, requiring further advancements in reliability estimation. Although the adopted definition of reliability in this chapter is replication based, the focus remains on discrete points in time rather than the continuous testing model mentioned by Haertel (2006). These types of assessment models, familiar to some as ongoing formative assessments, still require additional development in terms of reliability information across all previously collected data, the definition of what a replication consists of in a model continually adding additional assessment information, as well as what we are attempting to make inferences about (and when) in an ongoing data stream.

Additional considerations warrant attention, including the exploration of whether automated item generation and item cloning can accurately predict item parameters or characteristics to a sufficient extent for computing CSEMs prior to administering an assessment. Moreover, evaluating reliability around test takers using plausible values, imputed item scores, automated scoring, and scoring that incorporates additional information from item latencies or item clicks is essential. It is particularly important to define the relevant error components and determine the composition of a replication in these scenarios.

Additional ways to communicate what different levels of reliability and CSEM translate to in practice are needed. For example, under classification consistency/accuracy, informing users of the likely percentage of test takers being misclassified either at a cut score or overall helps a user interpret what a difference in classification consistency/accuracy might mean in practice. Similar ways of translating or illustrating the impact of different levels of reliability for a particular inference would be beneficial. Because settings differ widely in the type of decisions one is making, for example, whether to have a student do additional work with finding common denominators prior to moving on to subtraction involving mixed fractions versus licensing an applicant to practice in medicine or law, having one-size-fits-all requirements on the level of reliability needed is inappropriate. However, providing guidelines on presenting reliability information to enhance users' understanding of what a particular level of reliability means in a specific context would be helpful. In addition, more guidance on which types of coefficients on which metrics ought to be reported in different contexts would seem appropriate.

Cronbach stated in 2004,

> I am convinced that the standard error of measurement . . . is the most important single piece of information to report regarding an instrument, and not a coefficient. The standard error, which is a report on the uncertainty associated with each score, is easily understood not only by professional test interpreters but also by educators and other persons unschooled in statistical theory, and also to lay persons to whom scores are reported. (p. 413)

Cronbach's plea warrants serious attention.

Since the fourth edition of *Educational Measurement*, there has been relatively little literature published on the integration of IRT, GT, and CTT, despite Haertel (2006) and others encouraging it. Hopefully, this chapter is a meaningful response to that goal, though there is still plenty of room for additional progress. Also, while there has been some progress on integrating the concepts of reliability and validity, to echo Haertel (2006), much remains to be done.

ACKNOWLEDGMENTS

The authors express gratitude to Edward H. Haertel at Stanford University, Seonghoon Kim at Hanyang University, and Feifei Yei at RAND for their invaluable contributions as reviewers, providing excellent and thorough reviews. They also extend thanks to Robert L. Brennan at the University of Iowa and Stella Y. Kim at the University of North Carolina at Charlotte for their valuable comments on the initial draft. Special appreciation is given to Huan Liu and Kuo-Feng Chang, graduate students at the University of Iowa, for their assistance with the literature review and data analysis.

REFERENCES

Abdalla, W. (2019). *Detecting rater effects in trend scoring* [Doctoral dissertation, University of Iowa]. https://doi.org/10.17077/etd.qh0s-2ij2

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Anastasi, A., & Drake, J. D. (1954). An empirical comparison of certain techniques for estimating the reliability of speeded tests. *Educational and Psychological Measurement, 14*, 529–540.

Andersson, B., & Xin, T. (2018). Large sample confidence intervals for item response theory reliability coefficients. *Educational and Psychological Measurement, 78*, 32–45.

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika, 18*, 1–14.

Attali, Y. (2005). Reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement, 29*, 357–368.

Attali, Y. (2011). *Automated subscores for TOEFL iBT® independent essays* (ETS Research Report No. 11-39). ETS. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02275.x

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181–198). Routledge.

Attali, Y. (2015). Reliability-based feature weighting for automated essay scoring. *Applied Psychological Measurement, 39*, 303–313.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment, 4*(3), 1–29.

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*, 125–141.

Bacon, D. R., Sauer, P. L., & Young, M. (1995) Composite reliability in structural equations modeling. *Educational and Psychological Measurement, 55*, 394–406.

Beiser, D., Vu, M., & Gibbons, R. (2016). Test–retest reliability of a computerized adaptive depression screener. *Psychiatric Services, 67*, 1039–1041.

Berk, R. A. (1980). A consumers' guide to criterion-referenced test reliability. *Journal of Educational Measurement, 17*, 323–349.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431–444.

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*, 364–375.

Boyd, B., Lankford, H., Loeb, S., & Wyckoff, J. (2013). Measuring test measurement error: A general approach. *Journal of Educational and Behavioral Statistics, 38*, 629–663.

Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Lindenberger, U., & Hertzog, C. (2018). Precision, reliability, and effect size of slope variance in latent growth curve models: Implications for statistical power analysis. *Frontiers in Psychology, 9*, 294.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement, 32*, 385–396.

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). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20*, 6–18.

Brennan, R. L. (2001c). *Generalizability theory*. Springer-Verlag.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). American Council on Education/Praeger.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*, 1–21.

Brennan, R. L. (2022). Generalizability theory. In B. Clauser & M. B. Bunch (Eds.), *The history of educational measurement* (pp. 206–231). Routledge.

Brennan, R. L., & Kane, M. T. (1977a). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277–289.

Brennan, R. L., & Kane, M. T. (1977b). Signal/noise ratios for domain-referenced tests. *Psychometrika, 42*, 609–625.

Brennan, R. L., Kim, S. Y., & Lee, W. (2022). Extended multivariate generalizability theory with complex design structures. *Educational and Psychological Measurement, 82*, 617–642.

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.

Brennan, R. L., & Lee, W. (2006). *Some perspectives on KR-21* (CASMA Technical Note No. 2). Iowa City: University of Iowa.

Brennan, R. L., & Lee, W. (2008). Correcting for bias in single-administration decision consistency indices. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 15–24). Universal Academy Press.

Brennan, R. L., Yin, P., & Kane, M. T. (2003). Methodology for examining the reliability of group mean difference scores. *Journal of Educational Measurement, 40*, 207–230.

Breyer, F. J., & Lewis, C. (1994). *Pass–fail reliability for tests with cut scores: A simplified method* (ETS Research Report No. 94-39). ETS.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A review on sample size determination for Cronbach's alpha test: A simple guide for researchers. *The Malaysian Journal of Medical Sciences, 25*, 85–99. https://doi.org/10.21315/mjms2018.25.6.9

Burt, C. (1936). The analysis of examination marks. In P. Hartog & E. C. Rhodes (Eds.), *The marks of examiners* (pp. 245–314). The Macmillan Company.

Burt, C. (1955). Test reliability estimated by analysis of variance. *British Journal of Statistical Psychology, 8*, 103–118.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*, 221–248.

Cheng, Y., Yuan, K. H., & Liu, C. (2012). Comparison of reliability measures under factor analysis and item response theory. *Educational and Psychological Measurement, 72*(1), 52–67.

Choi, J., & Wilson, M. R. (2018). Modeling rater effects using a combination of generalizability theory and IRT. *Psychological Test and Assessment Modeling, 60*, 53–80.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391–418.

Cronbach, L. J., & Warrington, W. G. (1951). Time-limit tests: Estimating their reliability and degree of speeding. *Psychometrika, 16*, 167–188.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (ACT Research Report Series No. 97–4). ACT.

DeCarlo, L. T. (2010). *Studies of a latent class signal detection model for constructed response scoring II: Incomplete and hierarchical designs* (ETS Research Report No. RR-10-08). ETS.

DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*, 333–356.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*, 145–168.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13*, 354–378.

Dimitrov, D., Rumrill, P., Fitzgerald, S., & Hennessey, M. (1999). Reliability in rehabilitation measurement. *Work, 16*, 159–164.

Divgi, D. R. (1989). Estimating reliabilities of computerized adaptive tests. *Applied Psychological Measurement, 13*, 145–149.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407–424.

Ellis, J. L. (2013). A standard for test reliability in group research. *Behavior Research Methods, 45*, 16–24. https://doi.org/10.3758/s13428-012-0223-z.

Feldt, L. S. (1975). Estimation of the reliability of a test divided into two parts of unequal length. *Psychometrika, 40*, 557–561.

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. (1990). The sampling theory for the intraclass reliability coefficient. *Applied Measurement in Education, 3*, 361–367.

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

Feldt, L. S., & Qualls, A. L. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement, 33*, 141–156.

Feldt, L. S., & Qualls, A. L. (1998). Approximating scale-score standard error of measurement from the raw score standard error. *Applied Measurement in Education, 11*, 159–177.

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

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement, 31*, 525–543.

Fife, D. A., Mendoza, J. L., & Terry, R. (2012). The assessment of reliability under range restriction: A comparison of α, ω, and test–retest reliability for dichotomous data. *Educational and Psychological Measurement, 72*, 862–888.

Foltz, P. W., Streerer, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 68–88). Routledge.

Frank, B. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement, 62*, 254–263.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*, 72–91.

Gianinazzi, M. E., Rueegg, C. S., Zimmerman, K., Kuehni, C. E., & Michel, G. (2015). Intra-rater and inter-rater reliability of a medical record abstraction study on transition of care after childhood cancer. *PLoS One, 10*(5), 1–13. https://doi.org/10.1371/journal.pone.0124290

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423–436.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4–19.

Gilmer, J. S., & Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika, 48*, 99–111.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. *Educational and Psychological Measurement, 66*, 930–944.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347–360.

Green, S. B., & Yang, Y. (2008). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*, 155–167.

Gulliksen, H. (1950). The reliability of speeded tests. *Psychometrika, 15*, 259–269.

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment Research & Evaluation, 11*(6).

Guttman, L. A. (1945). A basis for analyzing test–retest reliability. *Psychometrika, 10,* 255–282.

Gwet, K. L. (2012). *Handbook of inter-rater reliability* (3rd ed.). Advanced Analytics.

Gwet, K. L. (2014). *The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics.

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

Haertel, E. H. (2013, March). *Reliability and validity of inferences about teachers based on student test scores*. The 14th William H. Angoff Memorial Lecture was presented at The National Press Club, Washington, DC. https://www.ets.org/Media/Research/pdf/PICANG14.pdf

Haley, S. M., Chafetz, R. S., & Tian, F. (2010). Validity and reliability of physical functioning computer adaptive tests for children with cerebral palsy. *Journal of Pediatric Orthopaedics, 30*, 71–75.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 80–123). The Johns Hopkins University Press.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). American Council on Education and Macmillan.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10,* 159–170.

Handcock, G. R., & An, J. (2018). Digital ITEMS module 2: Scale reliability in structural equation modeling. *Educational Measurement: Issues and Practice, 37,* 73–74.

Hanson, B. A. (1994). *An extension of the Lord–Wingersky algorithm to polytomous items* [Unpublished research note]. ACT.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27,* 345–359.

Henchy, A. M. (2013). *Review and evaluation of reliability generalization research.* [Doctoral dissertation, University of Kentucky]. https://uknowledge.uky.edu/edp_etds/5

Holland, D. F. (2015). *Reliability generalization: A systematic review and evaluation of meta-analytic methodology and reporting practice* [Doctoral dissertation, University of North Texas]. https://digital.library.unt.edu/ark:/67531/metadc822810

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6,* 153–160.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13,* 253–264.

Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch model. *Journal of Educational Statistics, 15,* 353–368.

Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments. *Applied Psychological Measurement, 10,* 175–186.

Jarjoura, D., & Brennan, R. L. (1982). A variance components model for measurement procedures associated with a table of specifications. *Applied Psychological Measurement, 6,* 161–171.

Jarjoura, D., & Brennan, R. L. (1983). Multivariate generalizability models for tests developed according to a table of specifications. In L. J. Fyans (Ed.), *New directions for testing and measurement: Generalizability theory: Inferences and practical applications* (No. 18, pp. 83–101). Jossey–Bass.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19,* 203–220.

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement, 55,* 635–664.

Johnson, S., & Johnson, R. (2009). *Conceptualising and interpreting reliability*. Office of Qualifications and Examinations Regulation.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Jöreskog, K. G., & Sörbom, D. (2018). *LISREL 10 for Windows* [Computer software]. Scientific Software International.

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

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.

Kane, M. T. (2017). *Measurement error and bias in value-added models* (ETS Research Report No. RR–17–25). ETS.

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4,* 105–126.

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education,* 17, 221–240.

Kano, Y., & Azuma, Y. (2003). Use of SEM programs to precisely measure scale reliability. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics.* Springer. https://doi.org/10.1007/978-4-431-66996-8_14

Karimi, L. (2015). *Model-based reliability and validity of measurement models using structural equation modeling* [Unpublishing doctoral dissertation, Swinburne University of Technology].

Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika, 22,* 29–41.

Kelley, T. L. (1947). *Fundamentals of statistics.* Harvard University Press.

Kendall, M. G., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed., Vol. 1). Macmillan.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58*, 587–599.

Kim, K. Y., & Lee, W. (2018). Confidence intervals for weighted composite scores under the compound binomial error model. *Journal of Educational Measurement, 55*, 152–172.

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika, 77*, 153–162.

Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review, 11*, 179–188.

Kim, S., & Livingston, S. A. (2017). *Accuracy of a classical test theory-based procedures for estimating the reliability of a multistage test* (ETS Research Report No. RR-17-02). ETS.

Kim, S. Y., & Lee, W. (2019). Classification consistency and accuracy for mixed-format tests. *Applied Measurement in Education, 32*, 97–115.

Kim, S. Y., & Lee, W. (2020). Classification consistency and accuracy with atypical score distributions. *Journal of Educational Measurement, 57*, 286–310.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. Springer.

Kolen, M. J., & Harris, D. J. (1987, April 20–24). *A multivariate test theory model based on item response theory and generalizability theory* [Paper presentation]. American Educational Research Association Annual Meeting, Washington, DC, United States.

Kolen, M. J., & Lee, W. (2011). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practice, 30*, 15–24.

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.

Kolen, M. J., Wang, T., & Lee, W. (2012). Conditional standard errors of measurement for composite scores using IRT. *International Journal of Testing, 12*, 1–20.

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.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika, 28*, 221–238.

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika, 39*, 491–499.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Lawrence, I. M. (1995) *Estimating reliability for tests composed of item sets* (ETS Research Report No. RR-95-18). ETS.

Lee, W. (2005). *Classification consistency under the compound multinomial model* (CASMA Research Report No. 13). Center for Advanced Studies in Measurement and Assessment, University of Iowa. http://www.education.uiowa.edu/casma

Lee, W. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement, 31*, 255–274.

Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47,* 1–17.

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.

Lee, W., Brennan, R. L., & Kolen, M. J. (2006). Interval estimation for true raw and scale scores under the binomial error model. *Journal of Educational and Behavioral Statistics, 31,* 261–281.

Lee, W., Brennan, R. L., & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement, 33,* 374–390.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26,* 412–432.

Lee, W., Harris, D. J., Liu, H., & Chang, K. (2025). *Empirical investigation of various reliability statistics* (CASMA Research Report No. 58). Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Lee, W., Kim, S. Y., Choi, J., & Kang, Y. (2020). IRT approaches to modeling scores on mixed-format tests. *Journal of Educational Measurement, 57,* 230–254.

Linacre, J. M. (1989). *Multi-faceted Rasch measurement*. MESA Press.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Houghton-Mifflin.

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement, 9,* 13–26.

Livingston, S. A. (2018). *Test reliability—basic concepts* (ETS Research Memorandum No. RM-18-01). ETS.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179–197.

Longford, N. T. (1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics, 19,* 171–200.

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

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement, 17,* 510–521.

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.* Lawrence Erlbaum Associates.

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

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

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Measurement in Education, 8*, 452–461.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity of examination scores. *Applied Measurement in Education, 3*, 331–345.

Madison, M. J. (2019). Reliability assessing growth with longitudinal diagnostic classification models. *Educational Measurement: Issues and Practice, 38*, 68–78.

Marcoulides, K. M. (2019). Reliability estimation in longitudinal studies using latent growth curve modeling. *Measurement: Interdisciplinary Research and Perspectives, 17*(2), 67–77. https://doi.org/10.1080/15366367.2018.1522169

Margolis, M. J., & Feinberg, R. A. (2020). *Integrating timing considerations to improve testing practices*. Routledge.

Mollenkopf, W. G. (1949). Variation of the standard error of measurement. *Psychometrika, 14*, 189–229.

Moses, T., & Kim, S. (2015). Methods for evaluating composite reliability, classification consistency, and classification accuracy for mixed-format licensure tests. *Applied Psychological Measurement, 39*, 314–329.

Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods, 23*(2), 351–362. https://doi.org/10.1037/met0000132

Nicewander, W. A., & Thomasson, G. L. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement, 23*, 239–247.

Oberski, D. L., & Satorra, A. (2013) Measurement error models with uncertainty about the error variance, *Structural Equation Modeling: A Multidisciplinary Journal, 20*, 409–428. https://doi.org/10.1080/10705511.2013.797820

Ogasawara, H. (2009). Stratified coefficients of reliability and their sampling behavior under nonnormality. *Behaviormetrika, 36*, 49–73.

Park, R., Kim, J., Chung, H., & Dodd, B. G. (2017). The development of MST test information for the prediction of test performance. *Educational and Psychological Measurement, 77*, 570–586.

Patterson, C. H. (1955). The interpretation of the standard error of measurement. *The Journal of Experimental Education, 23*, 247–252.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*, 341–384.

Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17*, 359–368.

Petscher, Y., Mitchell, A. M., & Foorman, B. R. (2015). Improving the reliability of student scores from speeded assessments: An illustration of conditional item response theory using a computer administered measure of vocabulary. *Reading and Writing, 28*, 31–56. https://doi.org/10.1007/s11145-014-9518-z PMC 5053774

Pugh, R. C., & Brunza, J. J. (1975). Effects of a confidence-weighted scoring system on measures of test reliability and validity. *Educational and Psychological Measurement, 35,* 73–78.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika, 30,* 39–56.

Raju, N. S. (1970). New formula for estimating total test reliability from parts of unequal length. *Proceedings of the 78th Annual Convention of the American Psychological Association, 5,* 143–144.

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika, 42,* 549–565.

Raudenbush, S. W., & Jean, M. (2012). *How should educators interpret value-added scores?* http://www.carnegieknowledgenetwork.org/wp-content/uploads/2012/10/CKN_2012-10_Raudenbush.pdf

Raykov, T., & Penev, S. (2010). Evaluation of reliability coefficients for two-level models via latent variable analysis. *Structural Equation Modeling, 17,* 629–641. https://doi.org/10.1080/10705511.2010.510052

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9,* 195–212. https://doi.org/10.1207/S15328007SEM0902_3

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment, 31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Rogosa, D., & Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational Statistics, 16,* 157–252.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20*(4), 335–343. https://doi.org/10.1111/j.1745-3984.1983.tb00211.x

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment Research & Evaluation, 7*(14).

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research & Evaluation, 10*(13).

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9,* 99–103.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion reliability: Implications for validation research. *Personnel Psychology, 55,* 807–825.

Samejima, F. (1993a). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika, 58,* 119–138.

Samejima, F. (1993b). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika, 58,* 195–209.

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

Sánchez-Meca, J., López-Pina, J. A., Rubio-Aparicio, M., Marín-Martínez, F., Núñez-Núñez, R. M., López-García, J. J., & López-López, J. A. (2019, May 30). *REGEMA: Guidelines for conducting and reporting reliability generalization meta-analyses*. Leibniz Institute for Psychology Information.

Schmitt, T. A., Sass, D. A., Sullivan, J. R., & Walker, C. M. (2010). A Monte Carlo simulation investigating the validity and reliability of ability estimation in item response theory with speeded computer adaptive tests. *International Journal of Testing, 10,* 230–261. https://doi.org/10.1080/15305058.2010.488098

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23,* 347–362.

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods, 24*(1), 70–91. https://www.researchgate.net/publication/327483449

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika, 66,* 79–97.

Seo, D. G., & Jung, S. (2018) A comparison of three empirical reliability estimates for computerized adaptive testing (CAT) using a medical licensing examination. *Frontiers in Psychology, 9,* 681. https://doi.org/10.3389/fpsyg.2018.00681

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educational Assessment, 20,* 46–65.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13,* 265–276.

Sun, A., Valiga, M. J., & Gao, X. (1997). Using generalizability theory to assess the reliability of student ratings of academic advising. *The Journal of Experimental Education, 65,* 367–379.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30,* 251–275.

Terry, L., & Kelly, K. (2012). Sample size planning for composite reliability coefficients: Accuracy in parameter estimation via narrow confidence intervals. *British Journal of Mathematical and Statistical Psychology, 65,* 371–401.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 159–184). Lawrence Erlbaum.

Thissen, D., & Orlando, M. (2001). Item response theory for items scores in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Lawrence Erlbaum.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19,* 39–49.

Thomas, R., & Zumbo, B. (2012). Difference scores from the point of view of reliability and repeated-measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement, 72,* 37–43.

Thompson, W. J., Clark, A. K., & Nash, B. (2019). Measuring the reliability of diagnostic mastery classifications at multiple levels of reporting. *Applied Measurement in Education, 32,* 298–309. https://doi.org/10.1080/08957347.2019.1660345

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

Tisak, J., & Tisak, M. S. (1996). Longitudinal models of reliability and validity: A latent curve approach. *Applied Psychological Measurement, 20*(3), 275–288. https://journals.sagepub.com/doi/abs/10.1177/014662169602000307

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16,* 8–14.

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalizations: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62,* 562–569.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46,* 247–272.

Verhelst, N., & Verstralen, H. (2001). IRT models for multiple raters. In A. Boomsma, M. van Duijn, & T. Snijders (Eds.), *Essays on item response theory* (pp. 89–108). Springer.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 165,* 22–29.

Wainer, H., Wang, X. A., Skorupski, W. P., & Bradlow, E. T. (2005). A Bayesian method for evaluating passing scores: The PPoP curve. *Journal of Educational Measurement, 42,* 272–281.

Walsh, P., Thornton, J., Asato, J., Walker, N., McCoy, G., Baal, J., Baal, J., Mendoza, N., & Banimahd, F. (2014). Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the emergency department. *PeerJ, 2,* e651. https://doi.org/10.7717/peerj.651

Wan, L., Brennan, R. L., & Lee, W. (2007). *Estimating classification consistency for complex assessments* (CASMA Research Report No. 22). University of Iowa.

Wang, C. (2014) Improving measurement precision of hierarchical latent traits using adaptive testing. *Journal of Educational and Behavioral Statistics, 39,* 452–477.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37,* 141–162.

Wang, Z., & Yao, L. (2013). *The effects of rater severity and rater distribution on examinees' ability estimation for constructed-response items* (ETS Research Report. No. RR-13-23). ETS.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427–450.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58,* 21–37.

Wood, S. W. (2020). Public perception and communication around automated essay scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 133–150). Chapman & Hall/CRC.

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

Woodruff, D. J., & Sawyer, R. L. (1989). Estimating measures of pass-fail reliability from parallel half-tests. *Applied Psychological Measurement, 13,* 33–43.

Woodruff, D. J., Traynor, A., Cui, Z., & Fang, Y. (2013). *A comparison of three methods for computing scale score conditional standard errors of measurement* (ACT Research Report Series 20137). ACT.

Woodruff, D. J., & Wu, Y.-F. (2012). *Statistical considerations in choosing a test reliability coefficient* (ACT Research Report Series 2012, 10). ACT. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2012-10.pdf

Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement, 36,* 602–624.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). American Council on Education/Praeger.

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data—which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology, 16,* 93. https://www.frontiersin.org/articles/10.3389/feduc.2017.00049/full

Zhang, Y., Breithaupt, K., Tessema, A., & Chuah, D. (2006, April 8–10). *Empirical vs. expected IRT-based reliability estimation in computerized multistage testing* [Paper presentation]. A National Council of Measurement in Education Annual Meeting, San Francisco, CA, United States.

NOTES

1. Coefficient alpha can also be derived under the assumption of randomly parallel forms. It can be shown that alpha is equal to the generalizability coefficient of a $p \times I$ design in GT.
2. Refer to Brennan and Lee (2006) for some other perspectives on KR21, especially its relation to absolute error variance.
3. General procedures are available for the UAO consisting of a mixture of infinite and finite universes—that is, a mixed effects model. However, a univariate mixed model is often viewed as a special case of a multivariate model, in which there is a random effects model within each condition of a fixed facet.