2

Classical Test Theory and Item Response Theory

Christine E. DeMars

CTT

Classical Test Theory (CTT) is fundamentally a test-level theory, not an item-level theory. The **true score** (T) is the foundation of CTT. Each time we test an examinee, we obtain an **observed score** (X). Imagine that we could erase the examinee's memory and restore her physiological state to exactly what it was at the beginning of the test and then test her again, resulting in another X. Repeat this an infinite number of times. T is defined as the expected value (mean) of the examinee's Xs: $T = \varepsilon(X)$. Note that in CTT, the **true** in true score does not refer to some objective inner property of the person, which is sometimes termed the Platonic true score (Lord & Novick, 1968; Sutcliffe, 1965). The CTT true score simply indicates the expected value of the observed score over repeated testings under the same conditions. Conceptually, many people would think of true score as an examinee's score under ideal testing conditions – what the examinee could do if motivated and not anxious or distracted. But that is **not** what true score means in CTT.

Assumptions

Several mathematical assumptions (axioms[1]) are needed to define CTT. These assumptions can be found in many sources (e.g., Allen & Yen, 1979; Crocker & Algina, 1986; Lord & Novick, 1968). A selection of the key assumptions is described here. In the following, the subscript p will indicate the person (examinee). A dot in the subscript indicates the mean over the elements (in the immediate context, persons) represented by the subscript. Subscripts will be dropped for all variances, covariances, and correlations because those terms must refer to persons, not means across persons. These assumptions

[1] Raykov & Marcoulides (2011, p. 121) discourage the use of the word *assumption* in this context because these assumptions are not falsifiable. I am using the term to mean axiom or postulate. Also, Raykov & Marcoulides note that many of the assumptions I have listed do not necessarily follow from the definition of X = T + E. I have chosen the axioms used to derive the most common formulas used in CTT. Different axioms will lead to different formulas.

The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, First Edition. Edited by Paul Irwing, Tom Booth, and David J. Hughes.

are defined in terms of the population values. Later in this chapter, sample estimates will be indicated by the symbol ^ over the appropriate index. Variance will be symbolized by σ^2, covariance by σ, and correlation by ρ, each followed by the variable in parentheses. Note that the symbol for correlation is the Greek rho (ρ), which unfortunately looks very similar to *p*.

1 $X_p = T_p + E_p$. X and T were defined previously as observed score and true score, respectively. E indicates random error. Errors are linearly uncorrelated with true score: $\rho(TE) = 0$.
2 The errors from two measurements are uncorrelated, or linearly independent. If E_1 and E_2 are the errors corresponding to two measurements for person *p*, $\rho(E_1E_2) = 0$. It follows that $\rho(T_1E_2) = 0$ and $\rho(T_2E_1) = 0$.
3 $\varepsilon(X_p) = T_p$. It follows that $\varepsilon(E_p) = 0$. It also follows at the group level that $\varepsilon(E_\bullet) = 0$.
   Although $\sigma^2(E_p)$ may differ for individuals, the mean error variance within persons equals the error variance across persons. Because the group error variance can be estimated accurately with just two measurements (or two subdivisions of a single measurement), it is often substituted as an estimate for each examinee's error variance (Lord & Novick, 1968, p. 155). Within a person, $\sigma(E)$is called the standard error of measurement (SEM).
4 Test reliability is defined as the squared correlation between the observed score and the true score: $\rho^2(XT)$. This is equivalent to the proportion of variance in the observed score due to true-score variance: $\rho^2(XT) = \frac{\sigma^2(T)}{\sigma^2(X)}$ (see Allen & Yen, 1979, Ch. 3, or Lord & Novick, 1968, Ch. 3, for derivations). Or equivalently $\rho^2(XT) = \frac{\sigma^2(X) - \sigma^2(E)}{\sigma^2(X)} = 1 - \frac{\sigma^2(E)}{\sigma^2(T) + \sigma^2(E)}$.
5 Parallel measurements are defined as measurements that have equal means, equal true-score variances, equal error variances, and equal correlations with the true-score. The observed score variances must also be equal because $\sigma^2(X) = \sigma^2(T) + \sigma^2(E)$. If X' is an observed score on a measurement parallel to X, $E(X_p) = E(X'_p)$, $\sigma^2(E) = \sigma^2(E')$, $\sigma^2(X) = \sigma^2(X')$, $\rho^2(XT) = \rho^2(X'T)$, and $\sigma(XX') = \sigma^2(T)$. It follows that $\rho(XX') = \rho^2(XT)$.

Tau-equivalent forms are defined as measurements that have equal true scores T_1 and T_2 for each examinee, but $\sigma^2(E_1)$ and $\sigma^2(E_2)$ may differ, where the subscripts 1 and 2 arbitrarily differentiate the two measurements. Essentially τ-equivalent forms are defined as $T_{2p} = T_{1p} + k$ for each examinee, where k is a constant. Thus, all parallel forms are necessarily τ-equivalent and all τ-equivalent forms are necessarily essentially τ-equivalent. The models are **nested**. The correlation between τ-equivalent scores is equal to the correlation between essentially τ-equivalent scores, because adding a constant to a score does not change the correlation. The covariance between essentially τ-equivalent scores equals the covariance between parallel scores; both equal $\sigma^2(T)$. But the correlation between parallel scores is not necessarily equal to the correlation between essentially τ-equivalent scores (unless the τ-equivalent scores happen to be parallel) because the denominator of the correlation coefficient includes $\sigma^2(E)$.

Congeneric forms may have different $\sigma^2(T)$ as well as different $\sigma^2(E)$ and different $E(X_p)$. $T_{2p} = cT_{1p} + k$, where c and k are constants. Essentially τ-equivalent scores are nested within congeneric scores. The assumption of linearly uncorrelated errors extends to τ-equivalent, essentially τ-equivalent, and congeneric scores.

Estimating reliability and the SEM

Hypothetically, the $\rho^2(XT)$ and $\sigma^2(E_p)$ could be known if we could test an infinite population, wipe out the memories of the participants and restore their psychological and physical states, and test them again and again. Then, we would know T_p for each person, as well as $\sigma^2(E)$, the mean $\sigma^2(E_p)$ and could calculate $\rho^2(XT) = 1 - \frac{\sigma^2(E)}{\sigma^2(T) + \sigma^2(E)}$. Or we could test each person in our infinite population just twice, as long as we did the memory erase and restore, and calculate reliability as the correlation between the two testings and then solve for $\sigma^2(E)$. In this hypothetical realm, the distinction between using the same test form or parallel forms is moot, because the properties of X and X' are identical.

Of course, this scenario is impossible. Instead, reliability is typically estimated either by administering the same test twice (test-retest reliability, coefficient of stability) or by administering two test forms that are as nearly parallel as possible (alternate forms reliability, coefficient of equivalence), or by treating parts within the test as separate measurements (internal consistency). With either of the first two designs, $\hat{\rho}^2(XT) = \hat{\rho}(XX')$ is the correlation between the two measurements. Then $\sigma^2(E)$ can be solved by substituting the empirical estimates $\hat{\sigma}^2(X)$ and $\hat{\rho}(XX')$ for $\sigma^2(X)$ and $\rho(XX')$ into a re-arranged form of the reliability definition:

$$\sigma^2(E) = \sigma(X)\sqrt{1 - \rho(XX')} \tag{2.1}$$

Or the error variance could be estimated as half the variance of the squared difference between the two scores: $\sigma^2(E) = \frac{\sum_p (d_p - d)^2}{2n}$, where $d_p = X_p - X'_p$. If we assume errors are normally and identically distributed, the SEM (square root of $\sigma^2(E)$) can be used to build confidence intervals around each X. Obviously conditional errors cannot be identically or normally distributed if X is the sum of the number of correct items (Lord & Novick, 1968, p. 502), but the approximation may be reasonable for scores away from the extremes of 0 and 100%. The assumption of normality is not necessary for the derivations in CTT; it only comes into play as one method of estimating confidence intervals.

Each of these methods is a lower-bound estimate of reliability if the assumption of linearly independent errors is met (Sitjsma, 2012; also see the discussion of these concepts in Lord & Novick, 1968, pp. 134–137). Alternate forms underestimates reliability to the extent that the forms are not truly parallel. Test-retest underestimates reliability to the extent that examinees' true scores change between measurements, possibly due to learning or even minor changes in mood or physical state. Some of the formulas based on subdividing the test, to be discussed in a later section, underestimate reliability to the extent that the parts are not essentially τ-equivalent. However, if the errors are correlated, reliability may be overestimated (Green & Yang, 2009; Komaroff, 1997; Zimmerman, Zumbo, & Lalonde, 1993), or underestimated less than it would otherwise be. In the test-retest design, for example, the assumption of

uncorrelated errors may be violated because examinees may remember specific items. For internal consistency estimates, if the part tests are individual items, items linked to a common context may have correlated errors.

Subdividing the test Sometimes it is not practical to administer two full-length test forms, either alternate forms or the same test on two occasions. If both measurements are taken at the same time, examinees may not be able to maintain concentration and motivation. If the measurements are obtained at separate times, the testing conditions may vary too much to be considered parallel; learning may have taken place in between the testings, or the examinees' psychological states may have changed beyond what the researcher is willing to consider error. Additionally, researchers generally want to minimize testing time. Testing takes time that could be spent in other ways, such as learning if the testing takes place in school, or a myriad of personal activities if the testing takes place outside of school. Thus, very often, reliability is estimated from a single measurement.

To estimate reliability from a single test form, X must be subdivided into two or more scores. The resulting reliability estimate is sometimes called a measure of internal consistency, because it is an index of how consistent scores from the parts of the test are. For example, to calculate a split-halves reliability estimate, the items are divided into two scores, X_1 and X_2, that are as nearly parallel as possible. The Pearson correlation between X_1 and X_2 is corrected by the Spearman–Brown prophecy formula $\rho(XX') = \frac{2\rho(X_1X_2)}{1+\rho(X_1X_2)}$, which estimates what the correlation would be if X_1 and X_2 were both full length. Or reliability can be estimated by Rulon's formula: $\rho(XX') = 2\left[1-\frac{\sigma^2(X_1)+\sigma^2(X_2)}{\sigma^2(X)}\right]$, where X is the full test score. If the test halves are parallel, these methods will be equivalent. Otherwise, the estimate from Rulon's formula will be slightly lower.

If X is subdivided into more than two parts, Rulon's formula generalizes to coefficient α (Cronbach, 1951), equivalent to Guttman's (1945) λ_3. This is often called Cronbach's alpha, although Cronbach discouraged this label and noted that the same result had been derived by others (Cronbach, 2004, p. 397). The computationally-simplest form of $\alpha = \frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma^2(x_i)}{\sigma^2(X)}\right)$, where *k* is the number of subtests and *i* indexes the subtest. Lowercase x is used here for the score on a subtest or item, to differentiate it from the total score X^2. As with Rulon's formula for split-halves, coefficient α is a lower-bound estimate of reliability. If the subparts of the test are not at least essentially τ-equivalent, coefficient α will underestimate the correlation between parallel measurements, the classic definition of reliability. However, as will be discussed in a later section, Cronbach (1951) broadened the definition of parallel forms to include randomly parallel forms. I will use the term classically parallel to indicate the traditional meaning

[2] This departs from the statistical convention of uppercase for random variables and lowercase for the value of a random variable.

of parallel as measurements with exactly the same T_p for each person and the same $\sigma^2(E)$, and randomly equivalent to indicate Cronbach's randomly parallel measurements. When the part tests are not τ-equivalent and the assumption of uncorrelated errors is met, coefficient α is an underestimate of the correlation between classically parallel forms, but it is an accurate estimate of the correlation between randomly equivalent forms (Cronbach, 2004, p. 204). Cronbach derived coefficient α for randomly parallel test forms; others (Guttman, 1945; Kuder & Richardson, 1937; Lord & Novick, 1968, pp. 89–90; Novick & Lewis, 1967) derived the same results for classically parallel forms, showing that coefficient α is a lower bound for classical reliability if the assumption of linearly uncorrelated errors is met.[3] Unless all of the items within the domain are parallel to each other, the X scores based on different sets of items drawn from the domain are unlikely to be parallel. The correlation between random forms is necessarily a lower bound to the correlation between parallel forms. Thus, if one wishes to estimate the correlation between random forms, instead of the correlation between parallel test scores, coefficient α is not an underestimate[4] (Cronbach, 2004, p. 400). However, it does not meet the classical definition of reliability.

At the extreme, each item can be considered a separate part of the test and coefficient α can be calculated based on the item scores. This is the most frequent way coefficient alpha is calculated. If the items are dichotomously scored, this reduces to the Kuder–Richardson formula 20 (KR-20, Kuder & Richardson, 1937). $KR-20 = \frac{k}{k-1}\frac{\sigma^2(X)-\sum_{i=1}^{k}P_i(1-P_i)}{\sigma^2(X)}$, where P_i is the proportion correct for item *i*. If all items are of equal difficulty, KR-20 is equivalent to $KR-21 = \frac{k}{k-1}\frac{\sigma^2(X)-kP.(1-P.)}{\sigma^2(X)}$, where $P.$ is the mean proportion correct. If KR-21 is applied when the items are not of equal difficulty, the corresponding SEM will contain the variance in item difficulty. Thus KR-21 ≤ KR-20.

Recall that coefficient alpha is a lower-bound estimate of reliability, equaling reliability only if the subdivisions of the test are at least essentially τ-equivalent. For long tests divided into just two or three parts, the scores may approach essential τ-equivalence. If there is a detailed test blueprint and the empirical characteristics of the items have been estimated, the test developer could split the items such that each subdivision of the test covers the blueprint proportionally and the parts have similar difficulty and variance. This seems far less plausible when the parts used in coefficient alpha are individual items. It becomes yet less plausible when the items are dichotomous. It is unlikely that dichotomous items could be essentially τ-equivalent unless they have the same mean, due to the relationship between mean and variance for dichotomous items. As Feldt (2002, p. 40) noted, if dichotomous items with different means were essentially τ-equivalent, all of the difference in variance would have to be due to differences in error variance. If this were the case, the covariances between items, which are not affected

[3] If the errors are correlated, such as would be likely when several items refer to the same scenario, coefficient α, and other estimates of reliability, may overestimate the correlation between parallel forms.

[4] Strictly speaking, the sample estimate of the covariance between random forms based on coefficient alpha, $\hat{\sigma}(X_1X_2) = \alpha\hat{\sigma}^2(X)$, is an unbiased estimate, but $\alpha = \hat{\sigma}(X_1X_2)/\hat{\sigma}^2(X)$ is biased for finite samples because the ratio of unbiased estimates is not in general unbiased except under select conditions. This bias is generally small.

by error variance, would be equal for all item pairs, which Feldt notes is contradicted by clear evidence of heterogeneity in item covariances within most tests.

Sometimes researchers explore whether dichotomous items are essentially τ-equivalent at the latent level instead of the observed level. Conceptually, this means that there is a latent continuous score x^* underlying the observed score x for an item. If x^* exceeds a threshold, which may vary by item, x = 1. Otherwise, x = 0. The non-linear relationship between x and x^* is typically modeled with either a probit or a logistic model. Because the x^*s are continuous, the squared correlation between x^* and T is not dependent on the mean of x^* and the latent x^*s may be essentially τ-equivalent without necessarily being parallel. This model can be tested through the methods of confirmatory factor analysis (CFA). But even if the latent x^*s are essentially τ-equivalent, it does not change the fact that the observed dichotomous item scores (the xs) cannot be essentially τ-equivalent unless they have the same mean. Coefficient α is estimated from the x's, not the x^*'s so it will be a lower-bound estimate of reliability unless the items are classically parallel. To provide an example of this concept, data were simulated for two test forms, each with 10 items (see Appendix, Code 1, for simulation details). On each test form, each x^* was correlated 0.6 with T and had a variance of 1, with means ranging from −1.35 to 1.35. Thus, the x^*s were τ-equivalent. For convenience, each x^* was normally distributed, although this is not an assumption of CTT. To approach the asymptotic true values, 10,000,000 examinees were simulated. The correlation between the parallel scores X_1^* and X_2^*, the sums of the 10 x^*s on each form, was 0.84909. Coefficient alpha was 0.849125 for X_1^*; the estimates differed at the fifth decimal because the samples were finite. The x^*s were τ-equivalent within each test form, so coefficient α was an accurate estimation of reliability.

Next, the x^*s were dichotomized. Any value of $x^* < -0.5$ was coded 0, and any value of $x^* >= -0.5$ was coded 1. Because the x^*s had different means (they were τ-equivalent, not parallel), this was equivalent to using a different threshold for each item. The resulting observed xs were no longer essentially τ-equivalent, even though the underlying x^*s were. The proportion correct on each item varied approximately from 0.20 to 0.97. The new X_1 and X_2, now integer values between 0 and 10, were still parallel. The correlation between the parallel measurements X_1 and X_2 was 0.68340. But coefficient α was slightly lower, at 0.665041 for X_1 and 0.665003 for X_2. Thus, if one is estimating the reliability of observed scores, it is the parallelness, or lack thereof, of the observed item scores, not the latent scores underlying the items, which determines whether coefficient alpha is an underestimate of reliability.

The extent to which coefficient α underestimates reliability due to using non-parallel items may be fairly small, as in the example before. For another example, Feldt (2002) estimated the bias in coefficient α, compared to a reliability formula derived for congeneric scores, for varying configurations of a test composed of 50 dichotomous items. Even the most extreme differences in true-score variance, not intended to be realistic, yielded a bias of −0.027. However, when the part tests were not items but subtests of different lengths, or raters using a rating scale, the bias was non-negligible. Similarly, very extreme violations of essential τ-equivalence for continuous item scores have demonstrated considerable negative bias on coefficient α (Komaroff, 1997; Zimmerman et al., 1993).

Another part-test estimate of reliability is McDonald's (1999) ω, estimated through CFA. For ω, the part tests are only assumed to be congeneric, not necessarily essentially

τ-equivalent. For tests with a single factor, $\omega = \frac{\left(\sum_i \lambda_i\right)^2}{\left(\sum_i \lambda_i\right)^2 + \sum_i \Psi_i^2}$, where λ_i is the unstandardized loading of item i or subtest i and Ψ_i^2 is the error variance for item i. If the items are essentially τ-equivalent, ω will equal coefficient α; otherwise, α is a lower bound to ω (assuming errors are conditionally independent).

Returning to our simulated dataset with τ-equivalent continuous item, each item had a loading of .6 and an error variance of .64; thus $\omega = \frac{(10(.6))^2}{(10(.6))^2 + 10(.64)} = .849057$, which closely approximates both the empirical correlation between the parallel forms and the empirical estimate of coefficient α. For the dichotomized items, the CFA must be run using a linear model, ignoring the fact that the items are dichotomous (for example MPlus code, please see the Appendix, Code 2). If the nonlinear model, which would be correct for other purposes, is run, the resulting ω will be an estimate of the reliability of the x^*s, not the observed item scores (see Appendix, Code 3, for example MPlus code). With the simulated data, using the nonlinear model resulted in loadings ranging from 0.599 to 0.601 and corresponding error variances from 0.359 to 0.361, producing the same ω as the continuous items, which is clearly an overestimate of the correlation of .683 between the parallel forms composed of dichotomous items. But for the linear model, the loadings ranged from .257 to .481 (unstandardized loadings ranged from .045 to .240). These loadings are smaller, reflecting the increased error due to dichotomizing the items, and more variable, reflecting the loss of essential τ-equivalence. The estimated $\omega = .676$, slightly larger than coefficient α.

Permissible replications T_p was defined earlier as $\varepsilon(X_p)$, where each X is a parallel measurement, and reliability was defined as $\rho(XX')$, where X and X' are classically parallel. However, a researcher may actually be more interested in the correlation between test scores that are NOT strictly parallel.

> The correlation between truly parallel measurements taken in such a way that the person's true score does not change between them is often called the coefficient of precision... For this coefficient, the only source contributing to error variance is the unreliability or imprecision of the measurement procedure. This is the variance ratio that would apply if a measurement were taken twice and if no practice, fatigue, memory, or other factor affected repeated measurement. In most practical situations, other sources of error variation affect the reliability of measurement, and hence the coefficient of precision is not the appropriate measure of reliability. However, this coefficient can be thought of as representing the extent to which test unreliability is due solely to inadequacies of the test form and testing procedure, rather than the extent to which it is due to changes in people over time and lack of equivalence of nearly parallel forms. (Lord & Novick, 1968, p. 134)

Cronbach (1951) argued that "It is very doubtful if testers have any practical need for a coefficient of precision" (p. 307). Cronbach focused instead on test forms composed of randomly selected items drawn from a hypothetical domain of items; he termed

measurements based on such test forms to be randomly parallel. Cronbach and colleagues (Cronbach, Rajaratnam, & Gleser, 1963) soon expanded this concept to include other random differences in the measurements, not just random selection of items. They developed generalizability theory methods to assess the contribution of different measurement facets, such as items, time, or observers.

Lord and Novick (1968, Ch. 8) termed the traditional true score the **specific** true score, and termed the expectation over randomly equivalent measurements the **generic** true score. They labeled the randomly equivalent measurements **nominally** parallel, and labeled the correlation between X and the generic true score the **generic** reliability. For the generic true score, one must define a **permissible replication**[5] (allowable observation). The definition of a replication of X defines how broad T is. For example, Thorndike (1951, p. 568) provided a table of factors that might impact score variance. He discussed how decisions about which factors contribute to true score and which to error variance impact data collection designs.

Consider the variable of human weight. To get truly parallel measurements, we would measure each person's weight on the same scale twice, without allowing restroom breaks or food/water ingestion; this would be a permissible replication. The correlation between these parallel measurements would be the classical reliability estimate, the coefficient of precision. This would be a useful summary of the consistency of the scale. But suppose a group of researchers wanted to get an estimate of each person's weight prior to a study of a weight-loss program. It would be more useful for this purpose to weigh each person several times throughout the day and average the results; the generic true score estimated here is the average weight across the day. Permissible replications are defined to require the same clothing and the same scale, but allow for fluctuations due to hydration and food intake. Each person would have a distribution of errors that would combine these fluctuations with inconsistency in the scale. The average error variance, together with the variance in weight over the group of study participants, could be used to estimate a reliability-like coefficient. This would clearly be a lower bound to the classical reliability because the error term is larger, just as coefficient alpha is a lower bound because of the additional error variance due to randomly sampling items. Permissible replications might also be broadened to include the use of different scales; this would again broaden the scope of the generic true score to include the mean across different scales as well as different weighing occasions. The error term expands to include the variance in the scales' accuracies as well as their consistencies and the humans' weight consistencies.

The strict definition of reliability in CTT defines X and X' as the same or parallel measures, which implies that nothing about the examinees or the tests changes between the measurements. Cronbach, as described, preferred to allow different test forms composed of random selections of items to be permissible replications. In many cases for achievement tests, we also want to generalize over occasions where the examinee's content knowledge has not changed, but concentration, motivation, and physical comfort may have changed slightly. These latter facets are allowed to vary randomly and the generic true score is the expected value of the measurements. For a psychological trait that is purported to be stable over lengthy time periods, observations many years apart would be permissible replications. The generic T would then be defined as the expected

[5] Lord & Novick reserved the use of replication for parallel measurements. I am using it more broadly here.

value over many years, perhaps a lifetime. Any time-related variances would become part of the error. Thus, in the generic conception of T, any facet that varies randomly is error. This broader definition of error yields lower reliability estimates – hence coefficient α as a lower bound to classical reliability.

Sometimes a characteristic may vary across measurements but consistently affect the scores of all examinees in the same way. This is sometimes called systematic error, a bit of a misnomer because CTT defines error as random. These types of errors are not detectable by correlational methods. If one test form is more difficult than the other for all examinees to the same extent, it will not change the correlation between the two scores. The variance in X is not changed by subtracting a constant from each score. If we were able to estimate each student's individual error variance from repeated testings, and some test forms were more difficult or easy, this error would be part of each student's error variance (called the absolute error variance in g theory). But when we estimate the error variance by substituting $\hat{\rho}(XX')$ and $\hat{\sigma}^2(X)$ into Equation 2.1, the systematic difficulty effect is not reflected in the error because it has no effect on either the correlation or the observed score variance. Similarly, any characteristics of the examinee that a researcher wishes to consider error must randomly vary across measurements or they become part of the true score. For example, if there is only one testing occasion and an examinee's motivation remains consistently low throughout the test, low motivation will be part of that examinee's true score, not error. As Thorndike (1951, p. 566) noted, all systematic factors are part of the true score.

In summary, before gathering data to estimate reliability, researchers must clarify which facets of the testing situation they consider part of the true score and which they consider error. The testing conditions must be designed so that facets considered error vary. The broader the pool of permissible replications, the more the generic reliability will underestimate the classical reliability. Thus, the generic reliability is a lower bound to the classical reliability. However, a broader definition of permissible replications may better match the score user's intentions than strictly parallel replications, even if achievable, would.

Additional considerations in reliability and the SEM

Increasing test length will generally increase the reliability of the test. Assuming that the additional test section will be parallel to the existing test, the error variance will increase linearly but the true-score variance will increase geometrically: $\sigma^2(E_X) = k\sigma^2(E_Y)$ and $\sigma^2(T_X) = k^2\sigma^2(T_Y)$, where Y is the original form, X is the lengthened or shortened form, and k is the multiplicative factor. Based on this relationship, the Spearman–Brown prophecy formula can be used to estimate the reliability of the lengthened test: $\rho(XX') = \frac{k\rho(YY')}{1+(k-1)\rho(YY')}$. Notice how increasing the test length has diminishing returns. For example, if a test with 20 items has a reliability of .8, decreasing the test to 10 items decreases the reliability to .67, but adding 10 items to a test length of 20 only increases the reliability from .80 to .86.

When a test is administered to a group with more heterogeneous scores, the reliability should increase but the SEM should remain the same[6]. The reliability in the new group

[6] If the SEM varies with the true score, as it would if the test score were the sum of dichotomous item scores, the SEM will likely vary by group if one group has mostly extreme scores. CTT does not account for this.

can be estimated by setting the SEMs of the two groups equal, resulting in: $\rho_{NN'} = 1 - \frac{\sigma_X^2(1-\rho_{XX'})}{\sigma_N^2}$, where $\rho_{NN'}$ is the reliability in the new group, $\rho_{xx'}$ is the reliability in the existing group, and σ_N^2 and σ_X^2 are the variances in the new and old groups.

Validity (defined narrowly as the correlation between X and an external score Y) is limited by the reliability of the measures; the validity coefficient is attenuated by measurement error. The correction for attenuation can be used to estimate what the correlation between two measures would be if there were no measurement error, the latent correlation. $\rho(T_X T_Y) = \frac{\rho(XY)}{\sqrt{\rho(XX')\rho(YY')}}$. Note that for this formula, the reliability estimates should be the classical reliability estimate, the correlation between truly parallel measures. If one substitutes the correlation between randomly parallel measures, such as coefficient alpha or the correlation between alternate forms that are not perfectly parallel, one will likely overcorrect to the extent that the reliability is underestimated (Lord & Novick, 1968, p. 138).

Item-level indices

Even though CTT is fundamentally a theory about test scores, we have seen that item scores play a role in the most commonly used estimates of reliability, coefficient α, λ_2, and ω, if the part tests used in the computations are individual items. The item scores obviously play a role in the total score X and its mean, variance, and error variance. Indices of item discrimination and difficulty are typically used to decide which items are useful for measurement and which should be discarded and replaced. Item difficulty is typically indexed by the mean score on the item. For dichotomous items, this is *P*, the proportion of examinees who answered the item correctly or endorsed the item. Sometimes this is called the *P*-value, although that term is easily confused with the *p*-value from statistical significance testing and will be avoided here. *P* is thus an index of item easiness or item facility, but it is historically called item difficulty. *P* is obviously dependent on the ability distribution. Because *P* is limited to the range 0–1, estimates of *P* from samples of examinees with different ability distributions will have a nonlinear relationship. To make the relationship linear, sometimes a continuous latent score x^* is conceptualized as underlying the observed score x. The threshold above which x = 1 is used as the index of item difficulty. The classical method of estimating the threshold requires assuming that x^* is normally distributed and there is no correct guessing. Note that this is not generally an assumption of CTT, but is an additional assumption made specifically for the purpose of estimating the threshold. The threshold is defined as the z-score corresponding to *P*. Angoff and Ford (1973) multiplied this value by 4 and added 13, labeling the resulting index Δ.

Generically, item discrimination is any index of how well the item separates examinees with high values of X from examinees with low values of X. Items with high discrimination will add to the reliability of X. The most common index of item discrimination is the correlation between the item score x_i and the total score X. If the item is dichotomous, this is termed the **point-biserial** correlation. It can be calculated either with the usual Pearson correlation formula or equivalently through a shortcut formula specifically for dichotomous items. Usually, the point-biserial correlation for item *i* is computed without including x_i in X, termed the corrected item–total correlation.

The point-biserial correlation is impacted by *P*. Because the variance for a dichotomous item = $P(1-P)$, items with *P* near 0.5 have the greatest potential for high point-biserial correlations. *P* depends on the ability distribution, so the point-biserial correlation varies across samples with differing ability distributions. In contrast, the biserial correlation does not depend on *P*: $\rho_{biserial} = \frac{\sigma(x_i)}{Y}\rho_{point-biserial}$, where Y is the Y ordinate of the standard normal function at the z-score corresponding to the proportion correct. The biserial correlation again invokes the concept of the latent item score x^*; it is an estimate of the correlation between x_i^* and X. It is greater than the point-biserial correlation between x_i and X, especially for items with very high or very low means. As was true of the estimate of the item threshold, the biserial correlation requires the additional assumption of normality of x_i^* and no correct guessing.

The biserial correlation is useful for comparing estimates from samples that differ in mean ability but not in variance. The point-biserial correlation is applicable for choosing test items if the mean ability of future testing samples will be roughly comparable to the current sample. If an item is very easy for a group of examinees, it cannot discriminate well in that group, regardless of how well the biserial indicates it would discriminate if it had not been dichotomized.

As an aside while discussing x^*, another related index is the tetrachoric correlation between pairs of dichotomous items. The tetrachoric correlation is the estimate of the correlation between the x^*'s for items *i* and *j*, as opposed to the phi correlation, which is the Pearson correlation between the observed xs. The tetrachoric correlation will be greater than the phi correlation, just as the biserial is greater than the point-biserial. There is no simple equation for estimating tetrachoric correlations (see Kirk, 1973 for one method). An additional note is that the models underlying the computation of thresholds and biserial and tetrachoric correlations generally do not take into account correct guessing (Lord & Novick, 1968, p. 153). Carroll (1945, 1987) developed methods of correcting Pearson, tetrachoric, and biserial correlations for guessing.

Another somewhat common index of item discrimination is the difference in P for the top 27% of examinees and the bottom 27% of examinees (Kelley, 1939). Although this index loses information through artificially categorizing X, it may be easier for teachers and content panels to understand.

Item discrimination can be calculated for each response option. It should be positive for the correct option, and negative for useful distractors. Negative values indicate that the probability of choosing the option decreases as X increases.

In summary, items that are more discriminating increase the reliability of the scores. Conditional on the latent relationship with the true score, middle-difficulty items will have higher observed discrimination indices, such as the point-biserial correlation.

Strong True-Score Theory

Strong true-score theory (STT) provides an alternative to CTT. The **strong** in STT refers to the assumptions, which are stronger than those of CTT. In STT, one has to assume a given distribution for the error variance, and often for the true-score variance as well. Here, we will limit the discussion of STT to contexts where X is the sum of the number of correct dichotomous items. STT, in some sense, is an item-level theory. In contrast, the score X in CTT is simply a score. It does not have to be the sum of item scores.

Following Lord and Novick's (1968, Ch. 23) notation, I will call the true score ζ instead of T. $\zeta_p = \varepsilon_i(x_{pi})$, where x_{pi} represents person p's score on item *i*. ζ may be estimated as X/I, where I is the number of items on a test form. In CTT, the hypothetical replications of X for person p are assumed to be parallel. In the discussion of coefficient α, the concept of randomly equivalent scores (Lord & Novick used the term nominally parallel scores) was briefly introduced. STT is applicable to randomly equivalent scores as well as classically parallel scores. The items for a test form are viewed as random draws from a hypothetical domain of items. If all of the items in the domain are equally difficult and have the same linear relationship with T, the resulting scores are classically parallel. In contrast to the T_p in CTT, ζ_p is person p's expected value over the domain, not just the set of items on one test form. ζ_p is often called the domain score, or the universe score in generalizability theory. Lord and Novick (1968) called ζ the generic true score, to contrast to the specific true score T.

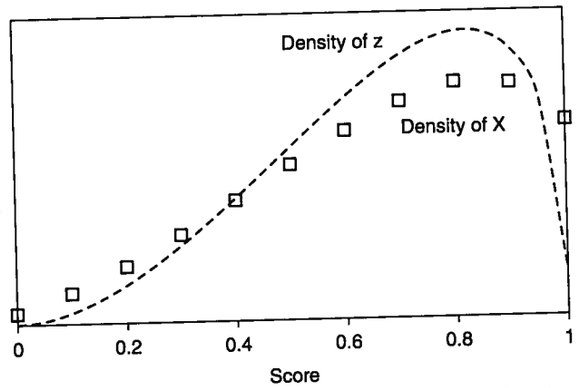
ζ is continuous but limited to the range of 0–1. One of the strong assumptions in STT is that, across people, ζ follows a specific distribution, typically either the two-parameter beta or the four-parameter beta.[7] The errors are assumed to have a binomial distribution. $X_p = \zeta_p I + E_p$, so the observed scores are said to have a beta-binomial distribution.

Using the binomial distribution, the SEM is calculated as $\hat{\sigma}(E_p) = \sqrt{\frac{X_p(k-X_p)}{k-1}}$, where k is the number of items. The binomial distribution produces heteroscedastic error variance. The SEM will be smaller for more extreme scores, and higher for X/k near 0.5. Estimating binomial standard errors allows test developers to meet Standard 2.14 (AERA/APA/NCME, 1999), which suggests reporting the SEM at various levels on the score scale. The $\hat{\sigma}^2(E)$ could be averaged across examinees and is equivalent to the $\hat{\sigma}^2(E)$ based on KR-21. The average error variance can be combined with $\hat{\sigma}^2(X)$ for a reliability estimate. When all items are of equal difficulty, this index will be equivalent to the correlation between parallel forms. If the items are not equally difficult, the binomial error variance contains additional variance due to differences in item difficulty. The resulting reliability-like index is an estimate of $\rho^2(X\zeta)$ where each examinee is randomly assigned a different form and the scores are not equated. It will thus tend to underestimate the correlation that would be observed if all examinees took the same set of randomly selected items. For the purpose of estimating this index, $\sigma^2(X)$ should be estimated from a sample in which examinees did take different sets of items. If $\sigma^2(X)$ is estimated based on a sample that took the same set of items, and the items differ in difficulty, it will not contain variance due to item sampling (Feldt, 1984), further decreasing the estimated ratio of $\frac{\sigma^2(X)-\sigma^2(E)}{\sigma^2(X)}$ because $\sigma^2(E)$ **does** contain variance due to item sampling. Or equivalently, one could estimate what $\sigma^2(X)$ would be if each examinee took a different sample of items by adding the variance in item difficulty to $\sigma^2(X)$.

For the ζs, the two-parameter beta distribution ranges from 0 to 1. The four-parameter beta distribution has a lower limit > 0 and an upper limit < 1. The four-parameter beta might better model scores where there is some correct guessing, such as multiple-choice item scores. However, estimating the parameters for the four-parameter beta from the observed data is considerably more complicated than estimating the parameters for the two-parameter beta (see Hanson, 1991; Lord, 1965 for details on the four-parameter beta). The parameters for the two-parameter beta distribution, often simply called the beta distribution without the qualifier **two-parameter**, can be estimated as:

$\hat{\alpha} = \left(-1 + \frac{1}{KR_{21}}\right)\hat{\mu}_X$, $\hat{\beta} = -\hat{\alpha} + \frac{I}{KR_{21}} - I$, where KR_{21} is the estimate of reliability using KR_{21}, $\hat{\mu}_X$ is the estimated mean score, and I is the number of items on the test form (Huyhn, 1979).

Combining the beta distribution for ζ and the binomial distribution for errors, one can estimate the X distribution. The resulting observed score distribution is applicable when either all items are of equal difficulty or each examinee responds to a different random subset of the items in the domain. To predict the X distribution for a specific test form with items of varying difficulty, a better estimate could be obtained from the compound binomial using subsets of items with similar difficulty (Lord & Novick, 1968, Ch. 23, section 10). Because the ζ distribution is continuous, it must be approximated at fixed points. At each point q on the ζ distribution, the distribution of X is binomial. The X distribution at q is then weighted by the density of ζ at q, then summed across the points. This approximates integrating the conditional X distribution over the ζ distribution. An example is shown in Figure 2.1. There were 10 items on the test, and alpha = 2.87, beta = 1.41 for the ζ distribution. Because the two-parameter beta-binomial distribution is equivalent to the negative hypergeometric distribution (Keats & Lord, 1962), the X distribution in this example could have been computed more directly, and without the need for approximating at discrete points on the ζ distribution, but the method described generalizes to any ζ distribution. This method could also be used to predict reliability for different populations by altering the density of ζ and re-weighting the distributions of X and E correspondingly.



Note: z has a continuous beta distribution and X has a discrete (beta-binomial) distribution.

Figure 2.1 Density of latent (ζ) and observed (X) scores.

[7] The beta distribution is typical for T when X is the sum of dichotomous scores. A Poisson distribution may be assumed for X when X is the number of words read correctly in a passage (Lord & Novick, 1968, Ch. 21; Rasch 1960/1980). Or for continuous scores, both errors and true scores might be assumed to be normally distributed.

Generalizability Theory

Generalizability theory (g theory) is an extension of CTT that allows for multiple facets of measurement. G theory is based on randomly equivalent (nominally parallel) measurements, not classically parallel measurements. As discussed earlier, often researchers want to include variance due to random differences in test forms or random differences in examinee factors on different days as error. Although the classical definition of parallel measurements does not allow for these differences, Cronbach's randomly parallel measurements or Lord and Novick's generic true score do. In these conceptualizations, error variance due to test form can be estimated by administering two test forms, or error variance due to occasion test-retest can be estimated by administering the same test twice. The error variance could include both if the test-retest data were collected with alternate test forms, but the two sources of error would be confounded. G theory allows for the separation of different sources of error, called facets. For example, the design might include 20 test items given on two occasions. The SEM would include error due to items and error due to occasion, and would include a separate estimate of each. Similarly, a design might include two raters scoring three writing samples for each examinee, and the SEM would include separate estimates of the error due to each facet. As in STT, the unit of analysis is x_p (a score from one item on one occasion, for example) instead of the total score X_p, and the x_p are often assumed to be randomly drawn from a larger domain (called the universe in g theory) of items, occasions, raters, and so on. However, the mathematical assumptions are the same as for CTT and only a single SEM is estimated for all persons.

After the error variance is estimated in g theory, reliability-like coefficients can be calculated, although Cronbach (2004) noted that often the SEM is more informative. The reliability-like indices are denoted as the G-coefficient and Φ-coefficient. The G-coefficient is used for comparing examinees who were scored based on the same random sample of items, raters, or whatever facets were included in the study. Anything about the sample of items (or raters, etc.) that impacts all examinees equally is not part of the error variance. In g theory, these are the main effects of the facet. Only the interaction between the examinee and the facet are considered error, called **relative** error because it is used when comparing examinees to other examinees who were exposed to the same random levels of the facets. This is similar to estimating reliability by correlations between two measurements; factors that add a constant to all examinee's scores do not impact the correlation or the observed score variance and thus are missing from the error variance if it is estimated as a function of the correlation and observed score variance. If there is only one facet, the G-coefficient is equivalent to coefficient alpha and the interaction between examinee and item is equivalent to the SEM estimated from Equation 2.1. Thus, the G-coefficient would be a lower bound to the correlation between parallel forms.[8] The G-coefficient is considered a reliability-like index, not an estimate of classical reliability but on a similar metric.

The Φ-coefficient is an alternative reliability-like index. The Φ-coefficient is useful for comparing examinees who were given different random draws of items or raters. The error variance used in calculating phi is called the **absolute** error variance. This error

[8] More precisely, it is also a lower-bound to the correlation between random forms because the ratio of the estimates is a biased estimate of the ratio unless certain conditions hold (Cronbach et al., 1963). But this bias is generally quite small.

variance, unlike the error variance for the G-coefficient or the error variance calculated in CTT by Equation 2.1, includes error variance systematically due to the difficulty of the items or harshness of the raters (main effects of the facets). If everyone received the same random set of items and raters, as often happens because the universe of items and raters is hypothetical, these main effects would add a constant to everyone's score. But if different examinees receive different random sets, as would occur if items were drawn from an itembank or subsets of raters from a pool were assigned to different examinees, these main effects should be counted as error because they randomly affect examinee's scores differentially. They would also be considered error for an individual, assuming the goal is to estimate the sampling distribution of the examinee's scores across random sets of measurement conditions, not just parallel forms. If the only facet in the design is items, and those items are dichotomous, the absolute error variance is equal to the average binomial error variance calculated in STT. Despite this equivalency, g theory was not derived for dichotomous scores, not does it require any assumptions about the distributions of true scores or errors.

This brief description of g theory was only enough to place it in the perspective of CTT. For an introduction to g theory, readers should consider Meyer (2010), or Shavelson and Webb (1991), followed by more advanced topics in Brennan (2001).

IRT

In contrast to CTT, IRT is an item-level theory. In IRT, a latent trait or ability or proficiency, symbolized θ, is posited to underlie the observed responses. The term **latent**, like the term **true** in true score, is not meant to imply an inner ability that would be elicited under ideal conditions. Latent simply means that we cannot measure θ directly, the way we can count the number of correct responses. IRT models assume a specific relationship between θ and the observed responses. IRT requires stronger assumptions about the response process and the error distribution.

Assumptions

1 Correct dimensionality: The item responses are assumed to be a function of the dimensions specified in the model. The most commonly used IRT models are unidimensional, so this assumption is often referred to as unidimensionality. Conceptually, the trait measured by the test may be a combination of several abilities, motivation, concentration, anxiety, and so on. If the same combination applies to every item, unidimensionality will hold. Unidimensionality can be assessed by many methods (see Hattie, 1984; Tate, 2003, for an overview).

2 Local independence: After controlling for the θ (or θs if the test is multidimensional) measured by the test, item responses should be locally independent. **Local** denotes conditional on θ, or controlling for θ. Globally (not conditional on θ), the item responses should not be independent or they would not be measuring anything in common; it is only locally that we assume the responses are independent. Strict local independence holds for the entire response string. When checking the assumption, however, researchers often test for local independence between pairs of items. This is termed **weak** local independence. See Kim, De Ayala, Ferdous, and Nering (2011) or Chen and Thissen (1997) for descriptions of tests for local independence.

3 The form of the model is correctly specified. Sometimes this assumption is stated as multiple assumptions, depending on the model selected. For example, some of the models described next assume there is no correct guessing. The one-parameter model assumes that all items are equally discriminating. Most models assume that the probability of scoring in the highest category increases as theta increases. All models assume that the relationship between theta and the item response is a continuous, defined function, often with either a logistic form or normal ogive form, which implies the errors are logistically or normally distributed. Assumptions about model form are generally checked by estimating the item parameters and checking the fit of the data to the model. For discussions of item fit, see Glas and Suárez Falcón (2003); Li and Wells (2006); Liang and Wells (2009); Orlando and Thissen (2000; 2003); Sinharay and Lu (2008); Stone (2000; 2003); and Wells and Bolt (2008). For discussions of model fit, see Kang and Cohen (2007); Kang, Cohen, and Sung (2009); Maydeu-Olivares, Cai, and Hernández (2011); and Whittaker, Chang, and Dodd (2012). Also see Chapters 8 and 9 of this volume.

Models

IRT models can be distinguished by whether they are unidimensional/multidimensional or dichotomous/polytomous. Dichotomous models are for items with only two response categories, such as right/wrong or agree/disagree, and polytomous models are for items with more than two response categories, such as an item scored with partial credit or a Likert type item. Here I will only describe dichotomous unidimensional models. Additional models are covered in Chapters 15 of Volume 1 and 16 of Volume 2. IRT models can be specified in logistic or normal (probit) forms. Normal models are mathematically more difficult to work with because calculating the response probabilities requires integration over the normal distribution. Logistic models are thus more commonly used. Sometimes a constant of 1.7 is placed in the logistic model so that the parameters will be nearly equivalent to those of the normal model. This is called the logistic model on a normal metric, and is the form I will use here.

The three-parameter logistic (3PL) model uses three parameters to model item responses.

$$P(\theta) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}} \qquad (2.2)$$

where $P(\theta)$ is the probability of correct response or item endorsement for an examinee with ability of θ_j, a_i is the item discrimination, b_i is the item difficulty, and c_i is the lower asymptote. An example of two 3PL functions is shown in Figure 2.2. The higher the value of the *a*-parameter, the more rapidly response probabilities increase as a function of theta. Thus, the *a*-parameter describes how well the item discriminates among theta levels. The higher the *b*-parameter, the less likely a respondent is to score 1. Unlike the CTT item difficulty index, a higher *b* indicates a more difficult item. The *c*-parameter is a lower asymptote indicating the probability of correct response for an examinee with very low θ. It is sometimes called the guessing parameter because one way that examinees of very low θ may choose the correct response is by guessing. It is also possible that some

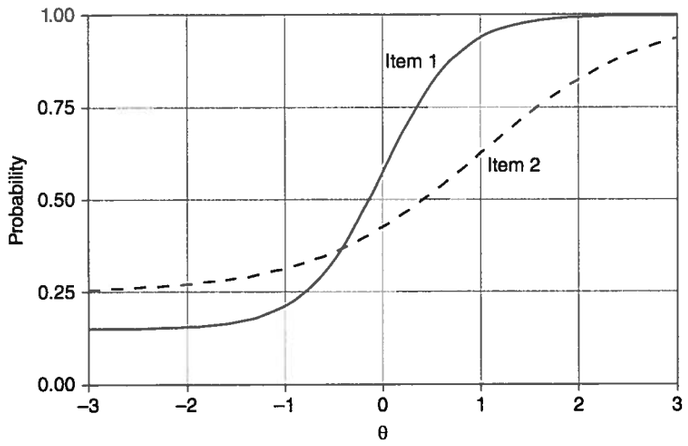


Figure 2.2 3PL items.

proportion of examinees know the information assessed by the item even if they are generally low on the θ measured by the test as a whole.

The phrase response probability, as used in this context, can be a bit ambiguous. Wainer, Bradlow, and Wang (2007, pp. 25–27) pointed out that, for most tests, it makes little sense to literally think about the response probability as the probability that examinee *j* will get item *i* correct. For example, based on the item parameters and θ_j, $P(x_{ij} = 1)$ might equal 0.5. But if we were to repeatedly test examinee *j* with this same item (with a memory erase, as in CTT), the examinee would likely select the same response very frequently, either because he knows the right answer or is consistently drawn to the same wrong answer. Instead, the response probability can be conceptualized in terms of a subpopulation of examinees with the same θ value. The response probability represents the probability that a randomly selected examinee from that subpopulation would answer the item correctly. Or it could be conceptualized in terms of a hypothetical population of items with the same parameters where the response probability represents the probability that the examinee would respond correctly to a randomly selected item with these parameters.

In the 2PL model, the *c*-parameter is fixed to zero and thus is removed from the model. In the 1PL model, a single *a*-parameter is estimated for all items and only the *b*-parameter differs. The 1PL model is equivalent to the Rasch model. The Rasch model was originally specified in terms of log-odds, not probabilities. It is now often specified similarly to the 1PL model, but without the 1.7 and using different symbols for the item difficulty and person ability/trait. The way the model is identified is also typically different in IRT and Rasch models. Two constraints need to be applied to fix the indeterminancy in the metric, one to center the scale and another to set the scale of the units. In IRT, the most common choice is to set the mean of the θs to 0 and the standard deviation of the θs to 1. Typically, Rasch modelers set the *a*-parameter to one, which frees the variance of θ, and the mean of the *b*s to 0, which frees the mean of θ. So the more discriminating the test, in terms of the IRT *a*-parameters, the greater the variance of the Rasch abilities. This illustrates that, like in CTT, the more heterogeneous the examinees' abilities are, the better we can discriminate among them.

If the model fits the data in all populations (an important caveat), the IRT item parameters are population invariant; they are the same in all populations. Aside from sampling error, the estimates of the parameters will vary by a linear transformation due to the indeterminacy in the metric. When the ability distributions of the populations differ, this cannot be true of the CTT item difficulty P; there is a nonlinear relationship between the item difficulty in different ability distributions. The point-biserial correlation, as discussed, depends on the item difficulty and also is not population invariant. However, if the ability distributions are not too discrepant, the relationships between estimates in different populations are often nearly linear (Fan, 1998). Additionally, if P is transformed to the threshold on the cumulative normal distribution corresponding to P, and the point-biserial correlation is transformed to the biserial correlation, invariance can be obtained if the additional assumptions of no correct guessing and normality of the underlying data are met.

The likelihood function

The likelihood function is an important concept in maximum-likelihood (ML) estimation. If the assumption of local independence is met, the likelihood of any observed response string, say 11110111001111000001000, is the product of the likelihoods of the individual item responses. For dichotomous items, the likelihood of an observed response of 1 is $P(\theta)$ and the probability of a response of 0 is $1 - P(\theta)$. If the item parameters are known, or treated as known, the likelihood is a function of θ. As long as there is at least one 0 and one 1 in the response string, the resulting function will have a maximum, and the value of θ where that maximum occurs is the ML estimate of θ. Figure 2.3 is an example of a likelihood function. Typically, the maximum is found using the Newton–Raphson method.

Sometimes information about the population distribution is used in estimating θ, employing Bayesian principles. The posterior distribution is the likelihood function multiplied by a prior distribution specifying the population density. The ML estimate

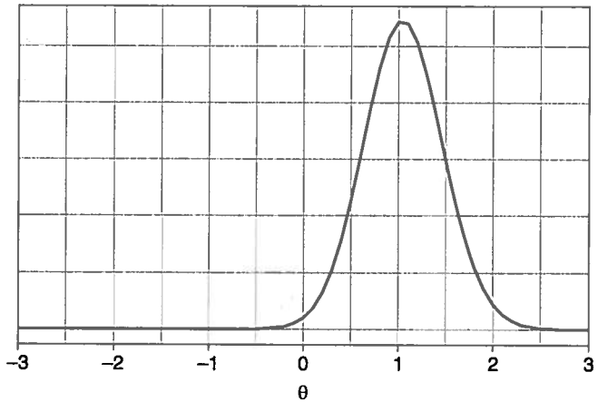


Figure 2.3 A likelihood function that reaches its maximum just above $\theta = 1$.

based on the posterior distribution is called the **modal-a-posterior** (MAP) estimate. $\hat{\theta}$ could instead be estimated as the mean, not the maximum of the posterior distribution, an estimate termed the **expected-a-posterior** (EAP, with the **expected** meaning **expected value**).

To estimate the item parameters, the $P(\theta)$ or $1 - P(\theta)$ in the likelihood is multiplied across people within an item. In most contexts, the individual θs are not known, and the likelihood is integrated across the θ distribution to reduce the dimensionality. This process is called maximum marginal likelihood (MML, also written as marginal maximum likelihood). The integration is approximated by the method of quadratures. The θ distribution may be assumed normal, or the shape of the distribution may be estimated as part of the MML process.

Item information and test information

Reliability and the SEM are major concepts in IRT, just as they are in CTT. Information underlies both concepts. Information is a statistical term, not unique to measurement. Information is the inverse of the asymptotic variance of a maximum-likelihood estimate. Thus, not only $\hat{\theta}$ but also each of the item parameters has an estimate of information and error variance. When the term **information** is used without a qualifier, it typically refers to the information for $\hat{\theta}$. The square root of the inverse of information ($\sqrt{\frac{1}{I(\theta)}}$) is the standard error of the parameter estimate. The standard error of $\hat{\theta}$ is the SEM, although some prefer to reserve the term SEM for observed scores. Because information is a function, not a single value, the standard error of $\hat{\theta}$ is also a function. An example is shown in Figure 2.4. In typical test forms, there is more information for middle values of θ, so the standard error is lower for these values. This contrasts with STT, where the standard errors are higher for those with observed scores nearer 50%. This difference can be explained by the difference in metrics. For examinees with very high or very low

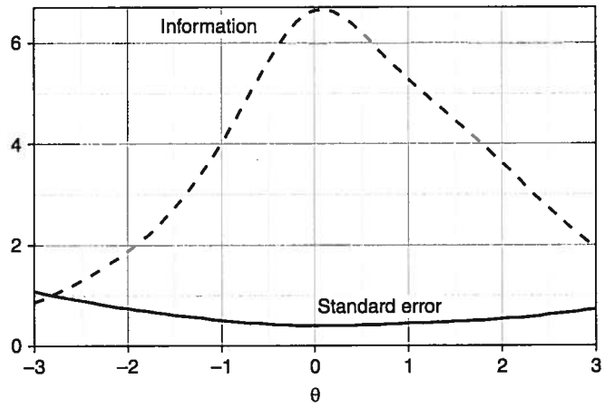


Figure 2.4 Information and standard error functions for θ.

θs, it can be difficult to estimate exactly how high or low θ is. But these examinees would consistently have observed scores near 100% or 0% on randomly selected test forms.

The information for θ can be decomposed into a series of item information functions. A useful property is that the item information functions sum to the test information function. Thus, the effect of adding or deleting an item or set of items can be easily calculated. In CTT, adding or deleting items changes T, so reliability and the SEM need to be re-calculated. If the IRT model fits, adding or deleting an item does not change the information function for any other items.[9] For dichotomous items, item information peaks at the difficulty value if the data follow a 1PL or 2PL model and somewhat above the difficulty value if the data follow a 3PL model. When item discrimination is high, the information is high at its peak but lower at other points. When item discrimination is low, information is spread over a broader theta range but is never very high. The higher the *c*-parameter, the lower the information, especially in the lower range of θ.

Items can be selected more precisely to provide information in specified ranges. Tests often are designed to have more information where the examinees are concentrated. If the intended examinee population matches the population used to center the metric, typically items would be chosen to provide the most information over a range between −2 and 2. If the test will be used on a more able group than the group used to center the metric, perhaps to select scholarship or award winners, more difficult items would be selected to provide more information in the top range. If the test has a passing standard, items might be selected to provide peak information at the cutscore, regardless of the θ distribution. Thus, the process of selecting items to fit an information function can be more systematic than the CTT method of selecting the more discriminating (usually middle-difficulty) items. Additionally, even if items have never been administered together on the same test form, once the items have been scaled to the same metric, possibly through a series of anchor items, the information for any set of items can be easily calculated, if the assumptions hold.

The information function is easier to work with than the standard error function because of its additive property. After the information function is calculated, the standard error function can be calculated. Due to the non-linear relationship between information and standard error, adding or deleting items has a non-linear effect on the standard error. Adding items has a decreasing impact on the standard error as the standard error gets smaller.

Information and the SEM are more useful than reliability because they are conditional on θ. But sometimes score users want a single summary of score precision. Reliability can be calculated from the CTT definition $\frac{\sigma^2(X)-\sigma^2(E)}{\sigma^2(X)}$, substituting the average standard error of θ for $\sigma^2(E)$ and the variance of the ML[10] $\hat{\theta}$s for the observed score variance, termed the empirical reliability (du Toit, 2003, p. 34). Or the theoretical reliability

[9] Of course, if the model does not fit, deleting a subset of items may change the estimation of $\hat{\theta}$ and all of the parameters of the other items.

[10] The formula is slightly different for Bayesian $\hat{\theta}$s because the variance of Bayesian estimates is the estimated true score variance.

can be estimated by integrating the information over the distribution of theta (either the hypothetical distribution or the distribution estimated along with the item parameters) and substituting the corresponding marginal standard error estimate. As in CTT, reliability must be in reference to a particular population. Reliability depends on the group variance.

One caveat regarding information: as noted in the introductory paragraph for this section, information is a property of ML estimates. In IRT, $\hat{\theta}$ is sometimes estimated by ML, with or without a Bayesian prior. The information function can be adjusted for the prior if the prior has a second derivative; for example, the second derivative of the normal distribution is the inverse of the standard deviation, so if a standard normal prior is used, a constant of 1 is added to the information across the θ range for the MAP estimates. EAP estimates are not ML estimates. The appropriate standard error for EAP estimates is the posterior standard deviation, which will be $<= \sqrt{\frac{1}{I(\theta)}}$ if I(θ) is based on the ML estimate without a prior but $>= \sqrt{\frac{1}{I(\theta)}}$ if I(θ) based on the ML estimates with a prior.

Chapter Summary

This chapter has provided an overview of CTT, STT as a variant of CTT with stronger assumptions and a focus on items instead of test scores, and g theory as an extension of CTT. With these theories, the true score or domain score or universe score is defined simply as the expected value of the observed score. Next, the focus moved to IRT. In IRT, a latent variable, θ, is posited to underlie the observed item responses. Although the observed responses are used in estimating $\hat{\theta}$, θ is not defined directly in terms of the responses the way T is defined in terms of X in CTT. Procedures for estimating the standard error and reliability of the scores were also discussed for the different score theories. In CTT and g theory, a single estimate of the SEM is obtained for all scores. In STT and IRT, the standard error is a function of ζ or θ. In STT, proportion-correct scores near 0 or 1 have the lowest standard errors. In IRT, $\hat{\theta}$s furthest from the item difficulties have the highest standard errors. In all test scoring theories, reliability, in contrast to the SEM, is group dependent.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Prospect Heights, IL: Waveland Press.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95–105.

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

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika, 10,* 1–19.

Carroll, J. B. (1987). Correcting point-biserial and biserial correlation coefficients for chance success. *Psychometrika, 47*, 359–360.
Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.
Crocker, L. C., & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Group.
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., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology, 16*, 137–163.
du Toit, M. (Ed.). (2003). *IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact.* Lincolnwood, IL: Scientific Software International, Inc.
Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357–381.
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. (2002). Estimating the internal consistency reliability of tests composed of testlets varying in length. *Applied Measurement in Education, 15*, 33–48.
Glas, C. A. W., & Suárez Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*, 87–106.
Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*, 121–135.
Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255–282.
Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes.* (ACT Research Report Series 91–5). Iowa City, IA: ACT.
Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49–78.
Huynh, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. *Journal of Educational Statistics, 4*, 231–246.
Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*, 331–358.
Kang, T., Cohen, A. S., & Sung, H. J. (2009). IRT model selection methods for polytomous items. *Applied Psychological Measurement, 33*, 499–518.
Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27*, 59–72.
Kelley, T. L. (1939). The selection of upper and lower grades for the validation of test items. *Journal of Educational Psychology, 30*, 17–24.
Kim, D., De Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement, 35*, 447–471.
Kirk, D. B. (1973). On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient. *Psychometrika, 38*, 259–268.
Komaroff, E. (1997). Effect of simultaneous violations of essential τ-equivalence and uncorrelated error on coefficient α. *Applied Psychological Measurement, 21*, 337–348.
Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.
Li, S., & Wells, C. S. (2006). *A model fit statistic for Samejima's graded response model.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 6–10.

Liang, T., & Wells, C. S. (2009). A model fit statistic for generalized partial credit model. *Educational and Psychological Measurement, 69*, 913–928.
Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30*, 239–270.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.
Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling, 18*, 333–356.
McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.
Meyer, P. (2010). *Reliability.* New York, NY: Oxford University Press.
Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1–13.
Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–64.
Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289–298.
Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded ed.). Chicago, IL: University of Chicago Press. First published 1960, Copenhagen: Danish Institute for Educational Research.
Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* New York, NY: Routledge.
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.
Sinharay, S., & Lu, Y. (2008). A further look at the correlation between item parameters and item fit statistics. *Journal of Educational Measurement, 45*, 1–15.
Sitjsma, K. (2012). Future of psychometrics: Ask what psychometric can do for psychology. *Psychometrika, 77*, 4–20.
Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37*, 58–75.
Stone, C. A. (2003). Empirical power and type I error rates for an IRT fit statistic that considers the precision of ability estimates. *Educational and Psychological Measurement, 63*, 566–583.
Sutcliffe, J. P. (1965). A probability model for errors of classification. *Psychometrika, 30*, 73–96.
Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159–203.
Thorndike, R. L. (1951). Reliability. In E. F. Lindquist & F. Everet (Eds.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.
Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* New York, NY: Cambridge University Press.
Wells, C. S., & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory. *Applied Measurement in Education, 21*, 22–40.
Whittaker, T. A., Chang, W., & Dodd, B. G. (2012). The performance of IRT model selection methods with mixed-format tests. *Applied Psychological Measurement, 36*, 159–180.
Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of reliability under violation of two assumptions. *Educational and Psychological Measurement, 53*, 33–49.

Code Appendix

Code 1 SAS latent and observed tau-equivalence.

```
options nocenter;
*illustrates difference between latent and observed tau-
equivalence;
%let seed=552954; *so can recreate later;
libname lib1 "C:\Christine";
data lib1.tau;
seed=552954;
do id=1 to 100000000;
  array a[10]; *continuous items;
  array b[10]; *another set of continuous items;
  array c[10]; *di items;
  array d[10]; *another set of di items;
  T=rannor(&seed);
  do i=1 to 10;
    a[i]=.6*T + .8*rannor(&seed)+(i-1)*.3 -1.35;
      b[i]=.6*T + .8*rannor(&seed)+(i-1)*.3 -1.35;
      if a[i]<-.5 then c[i]=0; else c[i]=1;
    if b[i]<-.5 then d[i]=0; else d[i]=1;
  end;
  Xstar1=sum(of a1-a10);
  Xstar2=sum(of b1-b10);
  X1=sum(of c1-c10);
  X2=sum(of d1-d10);
  output;
 end;
 run;
data _null_; set lib1.tau;
file "C:\Christine\delta.dat";
put c1-c10 d1-d10;
run;

 proc corr; var Xstar1 Xstar2; run;
 proc corr alpha nocorr; var a1-a10; run;
 proc corr alpha nocorr; var b1-b10; run;
  proc corr; var X1 X2; run;
 proc corr alpha nocorr; var c1-c10; run;
 proc corr alpha nocorr; var d1-d10; run;
```


Code 2 MPlus delta linear model.

```
TITLE: dichotomous items linear model
DATA: FILE IS "C:\Christine\delta.dat";
FORMAT is (10F2);
NOBSERVATIONS=10000000;
VARIABLE: NAMES ARE I1-I10;
ANALYSIS: ESTIMATOR=ML;
MODEL: f1 by I1-I10*;
f1@1;
[f1@0];
OUTPUT: STDYX;
SAVEDATA:
RESULTS are "C:\Christine\deltaLin.res";
```

Code 3 MPlus delta probit model.

```
TITLE: dichotomous items probit link model
DATA: FILE IS 'C:\Christine\delta.dat';
 FORMAT is (10F2);
 NOBSERVATIONS=10000000;
VARIABLE: NAMES ARE I1-I10;
   CATEGORICAL ARE I1-I10;
ANALYSIS: ESTIMATOR=WLSMV;
MODEL: f1 by I1-I10*;
   f1@1;
  [f1@0];
OUTPUT: STDYX;
SAVEDATA:
 RESULTS are "C:\Christine\deltaProbit.res";
```