Psychometrics Primer

If researchers analyze scores from psychological tests, such as self-report questionnaires about mood, observational measures of behavior completed by one or more raters, or group- or individually administered tests of cognitive abilities, they should know something about the statistical properties of test scores, or psychometrics. These properties include score reliability coefficients and validity coefficients. Reliability coefficients estimate score precision in a particular sample, and validity coefficients concern whether interpretations of test scores in a particular context of use are supported by relevant data. If scores are imprecise or have no trustworthy interpretation, then any results based on them may be meaningless. We begin with suggestions about how to select good psychological tests. Next, reporting standards for psychometrics by the American Psychological Association are summarized. Finally, basic psychometric concepts and statistics are described. Note that it is not possible to cover all aspects of psychometrics in a single chapter, but more advanced works are cited throughout, and they should be consulted for more information.[1]

SELECTING GOOD MEASURES AND REPORTING ABOUT THEM

It is just as critical in SEM as in other types of analyses to (1) select measures with strong psychometric properties and (2) report these characteristics in written summaries. This is because the product of measures, or scores, is what is analyzed. If the scores do not have good psychometrics, then the results can be meaningless.

Presented in Table P.1 is a checklist of descriptive, practical, and technical information that should be considered before selecting a measure. In the table, **test user qualifications** refers to the level of knowledge, experience, or professional credentials deemed neccessary for optimal test use. There are three basic levels (Urbina, 2014). The lowest level is “A,” which includes tests that require little, if any, special training to administer, score, and interpret. Level “B” designates tests that require specialized training in psychometrics and usually at least a Master’s-level degree in psychology, education, or a related field. The highest level, or “C,” designates tests strictly limited to those with very advanced training in psychometrics and assessment, such as at the Ph.D. level or its equivalent. Professional licensure or certification may also be required. Not all of these points may be relevant in a particular study, and some types of research have special measurement needs that may not be represented in the table. If so, just modify the checklist to better reflect a particular situation.

The *Mental Measurements Yearbook* (MMY), which as of late 2022 was in its 21st edition (Carlson, Geisinger, & Jonson, 2019) is a good source of information about commercial (i.e., copyrighted tests). The MMY includes both information about published tests and expert reviews (critiques) of test psychometrics, strengths, and weakness. It is available as a search-

[1] Parts of this presentation are adapted from Kline (2020).

TABLE P.1. Checklist for Evaluating Potential Measures

General

Test user qualifications (A, B, C)
Stated purpose of the measure
Attribute(s) claimed to be measured
Characteristics of samples in which measure was developed (e.g., normative sample)
Language of test materials
Costs (manuals, forms, software, etc.)
Limitations of the measure
Academic or professional affiliation(s) of author(s) consistent with test development
Publication date and publisher

Administration

Test length and testing time
Measurement method (e.g., self-report, interview, unobtrusive)
Response format (e.g., multiple choice, free response)
Availability of alternative forms (versions)
Individual or group administration
Paper-and-pencil or computer administration
Scoring method, requirements, and options
Materials or testing facilities needed (e.g., computer, quiet testing room)
Training requirements for test administrators or scorers (e.g., test user qualifications)
Accommodations for test takers with physical or sensory disabilities

Test documentation

Test manual available
Manual's description of how to correctly derive and interpret scores
Evidence for score reliability and characteristics of samples (e.g., reliability induction)
Evidence for score validity and characteristics of samples
Evidence for test fairness (e.g., lack of gender, race, or age bias)
Results of independent reviews of the measure

able electronic database in many university libraries. PsycTESTS by the APA is a searchable database and repository for noncommercial tests in several areas including psychology, psychiatry, mental health, education, neuropsychology, medicine, and social work, among others. Most entries include the actual test with links to relevant empirical studies.[1] Test information comes from peer reviewed journals, books, doctoral dissertations, and authors willing to share their tests with other researchers. Like the MMY, the PsycTESTS database is available in many university libraries. Maddox (2008) describes noncommercial measures in psychology, education, and business. Noncommercial measures are generally not protected by copyright, but as a professional courtesy you should ask the author's permission before using or adapting a particular test. The Educational Testing Service (ETS) Test Collection electronic database offers information about thousands of tests and research measures.[2] Some tests include instructions for administration or scoring. Tests are acquired from both publishers and researchers in the United States, Canada, the United Kingdom, and Australia. Finally, the interdisciplinary, peer-reviewed journal *Measurement Instruments for the Social Sciences* (MISS) publishes open access tests intended for research in psychology, political science, education, and other areas.[3]

Readers who have already taken a measurement course are at some advantage when it comes to select-

[1] *https://www.apa.org/pubs/databases/psyctests*

[2] *https://www.ets.org/test_link/about*

[3] *https://measurementinstrumentssocialscience.biomedcentral.com/about*

ing a test because they can critically evaluate candidate measures. They should also know how to evaluate whether those scores in their own samples are reliable and valid. Readers without this background are encouraged to fill in this gap. Formal coursework is not the only way to learn more about measurement. Just like learning about SEM, more informal ways to learn measurement theory include participation in seminars or workshops and self-study. A good undergraduate-level book that emphasizes classical measurement theory in psychology and education is Urbina (2014), and the graduate-level work by Furr (2022) deals with modern measurement theory.

Unfortunately, the state of practice about reporting on the psychometric characteristics of scores analyzed is too often poor. For example, Vacha-Haase and Thompson (2011) found that 55% of authors did not even mention score reliability in over 13,000 primary studies from a total of 47 meta-analyses of reliability generalization in the behavioral sciences. Authors mentioned reliability in about 16% of the studies, but they merely inducted values reported in other sources, such as test manuals. Such **reliability induction**, or inferring from particular coefficients calculated in other samples to a different population, requires explicit justification. But researchers rarely compare characteristics of their sample with those from cited studies of score reliability. For example, scores from a computer-based task of reaction time developed in samples of young adults may not be as precise for elderly adults. A better practice is for researchers to report estimates of score reliability from their own samples. They should also cite reliability coefficients reported in published sources (reliability induction) but with comment on similarities between samples described in those other sources and the researcher's sample.

Thompson and Vacha-Haase (2000) speculated that another cause of poor reporting practices is the apparently widespread but false belief that it is *tests* that are reliable or unreliable, not *scores* in a particular sample. In other words, if researchers believe that reliability, once established, is an immutable property of tests, then they may put little effort into estimating reliability in their own samples. They may also adopt a "black box" mentality that assumes that reliability can be established by others, such as a select few academics who conduct measurement-related research. The truth is that reliability and validity are attributes of scores in particular samples where the intended uses of those scores must also be considered.

Revised journal article reporting standards for quantitative studies by the APA (Appelbaum et al., 2018) emphasize complete and transparent reporting about psychometrics. Briefly summarized, these standards call on authors to (1) report values of score reliability coefficients for the scores analyzed (i.e., the researcher's sample), if possible. Examples include test-retest reliability coefficients in longitudinal studies, interrater reliability coefficients for subjectively scored measures, and internal consistency reliability coefficients for composite scales where total scores are summed over individual components or items. Also, (2) report the basic demographic characteristics of other samples if reporting reliability or validity coefficients from those sample(s), such as those described in test manuals or in the norming information about the test (e.g., reliability induction).

Measurement is a broad topic, so it is impossible to succinctly cover all its aspects, but familiarity with the issues considered next should help you to select good measures and report necessary information about scores generated from them. This presentation will also help you to better understand certain analysis options in CFA, the factor-analytic technique in SEM.

SCORE RELIABILITY

Score reliability is the degree to which scores in a particular sample are precise, or free from measurement error. *Precision* has a special meaning: If scores for the same participants maintain their *absolute positions* over variations in time (i.e., 2 or more occasions), test versions, item selections, or raters (for tests that are subjectively scored), the scores are precise. Thus, (1) persons with the highest scores in one variation will also tend to get obtain the highest score by the same margins in another variation, and (2) the corresponding pattern is true for persons with the lowest scores.

Reliability does *not* mean that each participant obtains exactly the same score over all conditions. Scores as just described would be perfectly precise, but so would any two sets of scores that preserve absolute differences among the participants. For example, in the two sets of scores for five cases, S_1 to S_5, respectively:

Set 1: 18, 23, 25, 29, 34
Set 2: 24, 29, 31, 35, 40

No scores for the same participant are the same across the two sets just listed, but absolute differences between any pair of scores in Set 1 is perfectly mirrored in Set 2, and vice versa. That the Pearson correlation $r = 1.0$ between the two sets of scores just listed describes the same characteristic. Some kinds of reliability coefficients are just Pearson correlations, which helps to make reliability analysis more familiar.

Reliability is estimated as one minus the proportion of total observed variance due to random error. These estimates are reliability coefficients, which for measure X are often designated with the general symbol r_{XX}. Because r_{XX} is a proportion of variance, its theoretical range is 0–1.0. For example, if $r_{XX} = .80$, then $1 - .80 = .20$, or 20% of total variance is unsystematic. But the remaining standardized variance, or 80%, may not all be systematic. This is because a particular type of reliability coefficient may estimate a *single source* of random error, and scores can be affected by multiple sources of error. As r_{XX} approaches zero, the scores are increasingly more like random numbers, and random numbers measure nothing. It can happen that an empirical reliability coefficient is less than zero. A negative reliability coefficient is interpreted as though its value were zero, but such a result ($r_{XX} < 0$) indicates a serious problem with the scores.

RELIABILITY METHODS AND COEFFICIENTS

Methods and coefficients in **classical measurement (test) theory**, which dates to the early 1900s (Jones & Thissen, 2007), are described next—later sections address two examples of more contemporary test theory. This discussion assumes that *speed* of performance has relatively little impact on test scores. This means that (1) if there is a time limit for participants, that limit is generous enough so that most examinees can complete the test. Also, (2) difficulty is manipulated by increasing or decreasing the complexity of the items, and the range of item difficulty is wide enough to accommodate examinees of different ability levels (Urbina, 2014). Such tests are called **power tests**. In contrast, difficulty for **speed tests** is determined by time limits so short that many, or perhaps most, examinees are unable to complete all items. Item difficulty in speed tests tend to be unform and relatively low; that is, each item is relatively easy, so performance speed, not knowledge, is what differentiates among the examinees. Speed tests require special methods to evaluate score reliability—see the classic work by Gulliksen (1950) for more information.

The most intuitive kind of reliability analysis is probably the **test-retest method**, where the same test is administered within the same sample but on two different occasions. When the scores are continuous, the **test-retest reliability coefficient** is just the Pearson correlation between the two sets of scores over time. If the two sets of scores are highly correlated, then error (i.e., change in absolute positions of scores over time) may be minimal. Thus, test-retest reliability coefficients measure **time sampling error**. For example, if the test-retest coefficient is .70, then at least $1 - .70$, or 30%, of the observed variation is due to time sampling error. It is critical to specify an appropriate retest interval, given the definition of the target concept. For example, a retest interval of 1 year may be appropriate when test scores are supposed to reflect enduring characteristics, such as general cognitive ability among adults. But a 1-year interval may be too long if the scores are expected to measure something less enduring, such as current mood. Exercise 1 asks you to interpret the meaning of a test-retest reliability coefficient that equals 1.0.

The **interrater reliability method** is for tests where scoring requires examiner judgment. An example is when examinee responses to oral questions are scored as correct (full pass), incorrect (full fail), or partially correct (i.e., in between full pass and fail) such that examiners must follow guidelines for assigning scores for each of the three outcomes just mentioned. The original test is administered once, but responses are independently scored by at least two different raters. When tests scores are continuous, such as a total score over the whole test, the **interrater reliability coefficient** is just the Pearson correlation between the two sets of scores. If the value of this coefficient is relatively high, then scores maintain their differences over raters. When test outcomes are categorical, such as the assignment of cases to mutually exclusive and nonoverlapping diagnostic categories, other coefficients can be calculated. An example is the **kappa coefficient**, which is a proportion of interrater agreement over cases corrected for chance agreement, or the likelihood that raters specify the same category by chance (Cohen, 1960).

The interrater reliability method estimates **rater sampling error**, or the extent to which score consistency is affected by the selection of a particular rater. Raters should be trained to correctly apply a scoring system, but then it is often necessary to periodically repeat the training or at least refresh raters familiarity with correct procedures. This requirement is due to **rater drift**, which is the tendency for raters to become more lax in their scoring over time. Retraining to avoid rater drift would be especially important in longitudinal studies where tests with subjective scoring systems are administered on multiple occasions to the same participants. Unfortunately, reporting about interrater reliability and the problem of rater drift is too often lacking in published studies. For example, Mulsant et al. (2002) reviewed a total of 63 published longitudinal randomized clinical trials where outcomes over time were assessed by raters. Values of interrater reliability coefficients, how the problem of rater drift was handled (if at all), or even the number of raters involved in the study were reported in less 25% of reviewed articles.

In the **alternate forms method**, two parallel versions of the same test are created, each of which is equally long and comprised of nonoverlapping sets of items selected from the same domain. For tests of ability or knowledge, it is also assumed that parallel versions are equally difficult. The availability of alternate versions of the same test helps to avoid practice effects in situations where participants are tested on two different occasions, such as before and after an intervention. Also, retesting is mandated in certain situations, such as when public school students who receive special education services are required to be reassessed within a prescribed amount of time. When test scores are continuous, the **alternate forms reliability coefficient** is the Pearson correlation between the scores across the two versions of the test. It measures **content sampling error**, or the extent to which absolute differences in scores over the two versions for the same participants are affected by chance selection of test content. For example, if the alternate forms coefficient is .80, then at least 20% of the observed variation is due to content sampling error.

A disadvantage of the alternate forms methods is that it requires at least two versions of a test, which may require twice the resources (or more) to developed compared with a single version. When just a single version of a test is available, a different method is needed, if content sampling error should be estimated. In the **split-half reliability method**, the whole and only version of a test is administered once, but then the test items are partitioned into two halves each with the same number of items (when the total number of items is an even number) or where one half has a single item more than the other half (when the total number of items is odd). Next, the total scores over all items for each half are computed for each participant and saved in the raw data file. The Pearson correlation between these sets of subtotal scores is r_{hh}, where the subscript indicates that each subtotal score is based on half the items in the original test. Finally, the value of r_{hh} is statistically corrected for the total items on the original test, and this adjusted result, r_{11}, is the **split-half reliability coefficient**. The correction is based on the Spearman–Brown prophecy formula for split-half reliability, or

$$r_{11} = \frac{2r_{hh}}{1 + r_{hh}} \tag{P.1}$$

where the constant "2" literally indicates that the whole test has twice as many items as each half of the test. For example, if $r_{hh} = .60$, then correcting for the number of items on the whole test, the split-half reliability coefficient is $r_{11} = .75$ after applying Equation P.1. In this example, at least 25% of the observed variation is due to content sampling error, which includes the particular method used to split the items of the original test into two equal sets.

A complication is that there are typically multiple ways to split a set of items into two halves. Examples include methods based on item position, such as when all even-numbered items (2, 4, etc.) are assigned to one half while all odd-numbered items (1, 3, etc.) are assigned to the half; that is, an odd–even split. Another position-based partition for, say, a 40-item test is that items 1–20 are assigned to one half while the remaining items, or 21–40, are assigned to the other half; that is, a first half–half, second–half split. There are still more possibilities, such as various random splits of the items from the original test into two halves. For example, if a test has, say, $N_i = 20$ items, then the number of possible splits is the one-half the combination of 20 items taken 10 at a time, or $.5 \times 20!/(10!)^2 = 92{,}378$, and each one has its own value of r_{11}, all of which could in theory be different values. Thus, (1) there is typically no single, unique value for a given test, and (2) variation in r_{11} values is a kind of sampling error, if no single method to split the items is clearly optimal. An exception is described next.

When test items become increasingly difficult, an odd–even split is probably optimal. This is because the two halves formed in this way should be approximately equal in difficulty. The split-half method applied with an odd–even split would be ideal when items increase in difficulty *and* there is a stopping rule for administration of the test, such as testing is discontinued when an examinee fails (i.e., the score is zero) five consecutive items. All remaining items, if any, would be not be administered, and their scores are assumed to all equal zero (i.e., they are considered failed). The rationale is that all remaining items are even more difficult than the failed items that corresponded to the stopping rule for a particular examinee, so missing data (unadministered items) are treated as though they were administered but then failed. Subtests on many individually-administered tests of general cognitive ability or scholastic achievement feature both increasingly difficult items and stopping rules. Exercise 2 asks you to calculate split-half reliability coefficients for a small dataset.

Content sampling error in the **internal consistency reliability method** is estimated at the level of individual test items; that is, whether participants' responses maintain their absolute differences over the whole set of items. Conceptually, an original test with N_i items is "split" into as many parts as items, or $1/N_i$ parts; that is, each item is essentially treated as a mini-test. Next, the degree of precision is estimated for each pair of items, of which there are a total of $N_i\ (N_i - 1)/2$ pairs. In a 20-item test, for example, there are 20(19)/2, or 190 different pairs of items. Finally, the average consistency over all items is calculated. If this average equals zero, then (1) there is zero response consistency over the whole set of items, and (2) the value of the **internal consistency reliability coefficient** also equals zero.

The most widely reported internal consistency coefficient—and also the most widely reported reliability coefficient of any kind (Thompson, 2003)—is **Cronbach's alpha**, also called the **alpha coefficient** or just plain **alpha** (Cronbach, 1951). When all test items are standardized (i.e., their scores are normal deviates where $M = 0$ and $SD = 1.0$), the equation for **standardized alpha** is

$$\alpha_S = \frac{N_i\,\overline{r_{ij}}}{1 + (N_i - 1)\,\overline{r_{ij}}} \tag{P.2}$$

where $\overline{r_{ij}}$ is the average Pearson correlation between all pairs of items. The whole ratio estimates the proportion of standardized variance in total scores computed over all N_i items (denominator) that is consistent, or shared between pairs of items (numerator). For example, given a mean interitem correlation of .30 for a set of 20 items, then

$$\alpha_S = \frac{20(.30)}{1 + (20 - 1)(.30)} = .90$$

That is, about 1 – .90, or 10% of total standardized variation is shared by the test's 20 items (i.e., it is consistent), but there are some important caveats to this interpretation that are considered momentarily.

A drawback of standardized alpha is that any differences in the item variances in the original (raw score) metric are lost when items are standardized. Another is that the analysis of standardized variables is not ideal when comparing results for the same measures over different samples (see the Regression Primer). For these reasons, the unstandardized form of the alpha coefficient, designated next as just "α" with no superscript, is generally preferred. The equation is

$$\alpha = \frac{N_i\,\overline{c_{ij}}}{\overline{s_i}^2 + (N_i - 1)\,\overline{c_{ij}}} \tag{P.3}$$

where $\overline{s_i}^2$ is the average variance over all items and $\overline{c_{ij}}$ is the average covariance for all pairs of items. The covariance for two continuous variables is the product of their Pearson correlation and standard deviations, or

$$c_{ij} = \text{cov}_{ij} = r_{ij}\,SD_i\,SD_j \tag{P.4}$$

which is an unstandardized measure of the linear relation between two variables that preserves the original (raw score) metrics of both variables. Because the covariance is an unstandardized statistic, its value has no fixed lower or upper limit for any pair of variables, unlike the Pearson correlation, which is a standardized statistic. But like the Pearson correlation, higher positive values of the covariance indicate greater consistency in the scores for the same cases. Likewise, $c_{ij} = 0$ means there is zero consistency. This because if $r_{ij} = 0$, then $c_{ij} = 0$, too (the scores over the variables have no linear relation). For exercise 3, you are asked to calculate and interpret the value of α for a set of raw data, and exercise 4 involves computing α from summary statistics for a set of $N_i = 3$ items.

There is a special relation between the value of α and those for the various split-half reliability coefficients, or r_{11}, that could be calculated for the same set of items: If items have equal variances, then the average of all possible split-half coefficients (e.g., odd–even split, half–half, second–half split, random split, and so on) equals that of α for the same variables. But the relation just stated may not hold if variances differ over items, which is more likely in real datasets than exactly equal variances. In general, the alpha coefficient is *smaller* than the average value of r_{11} to the degree that item variances are unequal. If there is no clearcut, optimal way to split the items into two halves, then the alpha coefficient may be preferred as a more general and deterministic (i.e., there is a single unique value) reliability coefficient than r_{11}. This is because the value of the alpha coefficient does not depend on the use of a particular method to split the items; that is, its value is more stable than that of r_{11} (Cortina, 1993). When the items increase in difficulty *and* there is a stopping rule such that not all items are not administered to every examinee, the alpha coefficient should *not* be calculated. This is because scoring missing responses for items not administered as failed responses (i.e., they are scored as zero) can distort the correlation matrix for the set of variables. Specifically, the value of alpha computed for such tests can result in artificially high values that are marginally less than 1.0 (Streiner, 2003). A better alternative in this case is r_{11} based on an odd–even item split.

There are additional potential complications in the interpretation of the alpha coefficient. The value of alpha (standardized or unstandardized) is affected by both the number of items (N_i) *and* the average consistency over pairs of items (respectively, $\overline{r}_{ij}$ or $\overline{c}_{ij}$). Specifically, the value of alpha generally increases as there are more items or the average correlation or covariance at the item level increases. This characteristic makes alpha more challenging to interpret. Suppose that $\overline{r}_{ij}$ = .02, which is virtually zero. For $N_i = 4$, $\alpha_S = .075$ (see Equation P.2), which is also practically zero. But for a much longer test, say, $N_i = 1{,}000$, then $\alpha_S = .953$, or nearly 1.0, but I think a researcher could hardly describe the longer test as "internally consistent," given the near-zero average intercorrelation correlation. In this second case, longer test length offsets low consistency over pairs of items because both it is basically the *product* of both factors just mentioned determine the value of alpha.

Here is a second example from Streiner (2003): Suppose that $\alpha_S = .95$ for a two-item test ($N_i = 2$). This is not a "good" result because $\alpha_S = .95$ implies that the items must be redundant, or so highly correlated that they measure nearly the same thing. In this example, $r_{12} = .905$, given the values of N_i and α_S for this example; exercise 5 asks you to verify this statement. Thus, responses to the two items are so highly correlated (close to 1.0) that one item or the other can be eliminated; that is, they are redundant as a set, and thus correspond to a single question asked of respondents, not two. Thus, sometimes the value of alpha can be *too high,* especially for a small number of items (Streiner, 2003). The characteristic that the value of alpha is determined by both test length and item-level consistency means that no single threshold or "golden rule" for a "good result" should automatically be applied to the alpha coefficient. Specifically, the widely used rule of thumb that $r_{XX} > .90$ indicates "excellent" score precision makes little sense for the alpha coefficient for the reasons just explained. Likewise, the heuristic that $r_{XX} = .70$ is a minimum "satisfactory" result for score precision also should *not* be blindly used with alpha. This is because having very many items can offset the effects of very low response consistency at the item level in the computation of alpha.

Equations P.3 and P.4 for alpha are based on variances, covariances, or correlations, which assume that items summed to form a total score are all continuous. But probably in most studies the alpha coefficient is computed for variables that are not continuous. Instead they are usually items with **Likert response scales**, where a relative small number of response options are represented with a set of equally-spaced integers, such as, "2" for "agree," "1" for "undecided," and "0" for "disagree." A problem is that the intervals between adjacent categories may not be actually equal on the underlying (latent) continuum that ranges from agree to disagree for this example. Thus, Likert response scales are usually seen as ordinal, not continuous. A second problem is that values of means and variances for items with Likert scales are generally arbitrary. This is because alternative numerical coding schemes, such as (10, 5, 0) for the same three responses for this example instead of (2, 1, 0), would work just as well as any other set of ascending or descending integers with equal intervals. Zumbo et al. (2007) described a version of alpha for ordinal variables such as Likert scale items, and how to analyze ordinal data in CFA is described in Chapter 18.

Summarized next are additional properties of the alpha coefficient, some of which can further complicate its interpretation—see Cortina (1993), Henson (2001), Streiner (2003), Tavakol and Dennick (2011), and Thompson (2003) for more information:

1. The alpha coefficient can be computed without performing a factor analysis with the items. This is because alpha already *assumes* a particular factor model, which is a one-factor model where all items were selected from a single common domain; that is, measurement is unidimensional.
2. Alpha further assumes **tau equivalence**, or that all items measure their common factor in the same way

(i.e., their unstandardized factor loadings all equal 1.0), but their error variances can be unequal (i.e., heterogeneity of error variance). This is a quite strict requirements that is probably violated in many, if most, data sets collected in real samples. Described in Chapter 14 are more flexible alternatives to the alpha coefficient with less stringent requirements, but they require the use of factor analysis.

3. Because the alpha coefficient assumes unidimensionality, obtaining a relatively high value of this reliability coefficient does *not* somehow prove or confirm this assumption. This is because it is possible to obtain relatively high values of alpha for factor models with ≥ 2 factors (i.e., multidimensional measurement), especially if factor correlations are relatively high.

4. As the number of items increases, the value of alpha can hide multidimensionality. With about 20 or more items, the value of alpha can be reasonably high, even though the items come from two unrelated domains. In general, it is better to directly test the hypothesis of unidimensionality in CFA by specifying and analyzing a single-factor model and computing an appropriate reliability coefficient for the model, which may not be alpha.

5. If items actually come from different domains—that is, a single-factor model is false and measurement is multidimensional—then alpha will underestimate the true precision of those items; that is, alpha is a **lower-bound estimate** of reliability in this case.

6. If items on a test are known a priori to be heterogeneous, such as when items are intentionally selected from different domains, then alpha should not be computed over all test items. Suppose that an engineering aptitude test has two types of items: Text-based problems and those that require the interpretation of data graphics. Responses over items from the two different domains just stated may be less consistent compared with responses within each set of items. Here, it would be better to calculate alpha separately within each set of items, text versus graphical, than for all items together.

7. Values of the alpha coefficient—and all other types of reliability coefficients, too—are subject to sampling error; that is, their values will change from sample to sample for the same test. Fan and Thompson (2001) described how to calculate confidence intervals based on alpha and other reliability coefficients. Such intervals explicitly define expected margins of error, given the sample size (i.e., more error is expected with smaller samples, and vice versa).

8. As proportions of variance, theoretical values of reliability coefficients, including alpha, range from 0 to 1.0. In practice, though, values of alpha can be negative (< 0). This can happen whenever the average intercorrelation correlation is negative, which also means that the average covariance will be negative, too. Negative empirical values of alphas are generally interpreted as though they were equal to zero, but $\alpha < 0$ indicates a problem with the scores. In general, all pairwise item correlations should be positive; otherwise, a mix of positive and negative correlations lowers internal consistency reliability.

Among items all drawn from the same domain, average interitem correlations less than zero can arise due to item wording or, specifically, a mix of items where some are positively worded but others are negatively worded. Consider the three items listed next that share the same Likert response categories (0 = *disagree*, 1 = *undecided*, 2 = *agree*):

1. My overall health is good.
2. I often feel unhealthy.
3. I worry little about my health.

Respondents who agree with the first and third positively worded items just listed will tend to disagree with the second items, which is negatively worded, and vice versa. These patterns would lead to negative correlations between the responses for the first and second items and also between the responses for the second and third item. The average intercorrelation over all three items could be less than zero, too. To avoid this problem, the researcher can use **reverse scoring** or **reverse coding** where scores for items that are negatively worded are reversed in a positive direction, to match the scoring positively worded items, or vice versa. The result is that negative correlations between positively worded items and negatively worded items are converted to positive correlations. In this example, scores of 0 for the second question (*disagree*) are converted to a score of 2, and scores of 2 (agree) for the same question are transformed to a score of 0. After these transformations for all respondents, the highest score on item 2 reflects better health, and vice-versa, just like for items 1 and 3. Now all three pairwise item correlations should be greater than zero.

To summarize, the alpha coefficient is widely reported in the literature, but its interpretation is rather challenging, especially if the researcher ignores its assumptions, which are quite demanding and, thus, generally unrealistic in real datasets. McNeish (2018) describes alternatives to Cronbach's alpha, including coefficients that can be calculated in CFA. Some of these CFA-based alternatives are described in Chapter 14.

FACTORS THAT AFFECT RELIABILITY

It probably does not surprise you at this point to learn that score reliability is affected by characteristics of the test. To summarize, tests that are longer (i.e., they have more items) tend to generate scores that are more precise than tests with fewer items. For example, in the split-half reliability method, the correlation between the scores from the two halves of the test, r_{hh}, must be corrected for test length (i.e., the whole test is twice as long as each half), and the resulting split-half reliability coefficient, r_{11}, is greater in value than r_{hh} except when $r_{hh} = 0$ (Equation P.1). Test length also affects the value of the alpha coefficient: Both the number of items and their variances and covariances determine alpha (Equation P.3).

You can think of longer tests as providing more information than shorter tests, *assuming that test items are of good quality*. Good items are well written and clear in their intended meaning, relevant to assessment of the target construct(s), of appropriate difficulty for tests of ability or knowledge, and have favorable values of item statistics (e.g., they correlate positively with other items from the same domain). Note that adding bad items to a test can actually *decrease* score reliability just as removing problematic items can *increase* reliability—see Urbina (2014, chap. 6) for more information. Heterogeneity of item content generally lowers internal consistency reliability, but the alpha coefficient assumes (i.e., requires) that all items were selected from the same domain. The scoring system for a test can also affect reliability: There is potentially no scoring error for tests with items that can be objectively scored such as math items for which there is a single correct answer and there is no credit for partially correct responses (i.e., scoring is pass–fail).

Score reliability is also affected by non-test factors that include characteristics of examiners (i.e., those who administer the test), test settings, and samples (i.e., examinees or respondents). Examiners should be properly trained in both test administration and scoring, especially if scoring is not completely objective. Periodic retraining in correct test administration or scoring may be needed to prevent rater drift over time. Examiners should also have the appropriate academic background, degree, or type of professional license to administer a particular test (i.e., user qualifications; Table P.1). The setting where tests are administered is another factor. For example, individually-administered tests of general cognitive ability, such as IQ tests, should be administered in rooms that are reasonably quiet and free from interruptions; otherwise, examinee scores may not be very precise.

Values of reliability coefficients are affected by sampling, that is, they vary over different samples all drawn from the same population. In this way, reliability coefficients are like basically all sample statistics in that their values are subject to sampling error. Score reliability can also vary widely as a function of examinee age, gender, ethnicity, level of education, or income, among other variables. Reliability coefficients are generally higher in heterogenous samples with greater ranges of individual differences. As samples become more homogeneous, such as due to range restriction, score reliabilities tend to decrease in value. Lack of motivation to participate in testing can also lower score reliabilities. For example, rates of random responding among participants recruited to anonymously complete questionnaires are surprisingly high, up to 30% or more in some samples (Osborne, 2013). Random responding can be motivated by apathy, fatigue, or intentional carelessness, among other reasons for being uncooperative.

CONSEQUENCES OF LOW RELIABILITY

Low score reliability generally has several deleterious effects in statistical analysis. Low reliability generally reduces the power of statistical significance tests. This means that ever larger sample sizes are needed to attain the same target level of power (e.g., at least .90) as score reliabilities decrease. Low reliability also generally decreases absolute effect sizes when dependent variables are measured with error. But if scores on both predictor and outcome variables in standard regression analysis are imprecise, then it can be more difficult to anticipate the exact pattern of consequences. This is because unreliability in predictors and criteria can artificially increase or decrease absolute values of regression coefficients (see the Regression Primer). This is especially true in the presence of correlated measurement error, or sources of imprecision that are shared by two or more variables (Williams et al., 2013). Fortunately, there are ways in SEM to control for measurement error that are generally unavailable in more standard statistical techniques, such as multiple regression and the analysis of variance (ANOVA). These issues are addressed in more detail in Chapter 15.

CLASSICAL AND MODERN VIEWS OF VALIDITY

In the classical model of validity that dates to the 1920s–1950s (e.g., Cronbach & Meehl, 1955), **validity** was broadly defined as whether a test actually measures what it was constructed to measure. Accordingly, the apex concept in this view is that of **construct validity**, which concerns whether test scores, or the observed data, can be interpreted as measuring the target hypothetical constructs, or latent variables, and *how well* the scores do so (Urbina, 2014). There is no single, definitive test of construct validity, nor is it established in a single study. Instead, measurement-based research usually concerns a particular aspect of construct validity. For instance, **criterion-related validity**, also called **external validity**, concerns whether test scores (X) relate to a criterion (Y) against which the scores can be evaluated. Specifically, are sample values of r_{XY} large enough to support the claim that a test explains an appreciable amount of the variability in the criterion? Whether an admissions test for graduate school predicts eventual program completion is a question of criterion-related validity.

Convergent validity and discriminant validity involve the evaluation of measures against each other instead of against an external standard. Variables presumed to measure the *same* construct show **convergent validity** if their intercorrelations are appreciable in magnitude. But if measures that supposedly reflect the same construct also share the same measurement method, their intercorrelations could be inflated by **common method variance**. Thus, the best case for convergent validity occurs when measures of the same presumed trait are each based on a different measurement method (Campbell & Fiske, 1959). Likewise, **discriminant validity** is supported if the intercorrelations among a set of variables presumed to measure *different* constructs are not too high, but this evidence is stronger when the measures are not based on the same method. If $r_{XY} = .90$ and these two variables are each based on a different measurement method, one cannot claim that X and Y assess distinct constructs. Hypotheses about convergent and discriminant validity are routinely tested in CFA.

Content validity deals with whether test items are representative of the domain(s) they are supposed to measure. Content validity is often critical for scholastic achievement measures, such as tests that should assess specific skills at a particular grade level (e.g., Grade 4 math). It is important for other kinds of tests, too, such as symptom rating scales. The items of a depression rating scale, for example, should represent the symptom areas thought to reflect clinical depression.

As in other kinds of statistical methods, SEM requires the analysis of scores with good evidence for validity. Because score reliability is generally required for score validity—but does not guarantee it—this requirement includes good score reliability, too (see Little et al., 1999, for exceptions). Otherwise, the accuracy of the interpretation of the results is doubtful. So using SEM does not free researchers from having to think about measurement (just the opposite is true).

Classical views of validity were limited by these features described by Urbina (2014): Validity was viewed as a property of tests, not of scores from tests in a particular context of use. Just as reliability is not an absolute, immutable property of a test, the same is true about validity. In the classical view, to be valid, scores from a test should measure some purported construct, but its definition reflected the test's author(s) understanding of that construct, which could be relatively idiosyncratic. Classical views of validity did not extend to situations where a test is used on a strictly empirical or practical basis, such as when predicting an external criterion from a set of test scores with little, if any, basis in theory about what the test measures. Finally, the idea of construct validity may not generalize to the study of complex multidimensional or theoretical constructs for which there are no clearcut, consensus definitions (e.g., intelligence).

More contemporary definitions of validity correspond to an approach by Kane (2013) described as **interpretation-use arguments**. This perspective concerns the plausibility and appropriateness of both the interpretation and the proposed uses of scores. That is, validity is not a fixed property of tests; rather, it involves the proposed interpretation and intended uses of the scores. As the range of potential generalizations from test scores increases, such as from an observed sample of performances (test data) to predicted performances in other settings, more evidence is needed. Thus, the definition of validity is not restricted to the question about what particular construct is measured by test scores. Instead, validity in the interpretation-use arguments perspective concerns the match between a proposed interpretation of test scores and the nature or scope of the evidence required to support that interpretation. For scores from the same test, different evidence may be required for any alternative interpretation. Additional concerns about validity were articulated by Messick (1995), who emphasized the qualities of relevance, utility, value implications, social justice and equality, and social consequences of test use and interpretation in validation. An example of the social consequences of testing includes the fair and accurate assessment of cognitive abilities among minority children. Tierney (2016) described the concept of **fairness** in assessment.

The emphasis on interpretation-use arguments is gradually replacing the idea of construct validity as the central organizing principle. Described in the standards for educational and psychological testing by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (2014)**.** are the four categories of validity evidence outlined next:

1. **Test content evidence** is especially relevant when scores should reflect knowledge, skills, or status in a target domain. An example is **mastery testing**, also called **domain-referenced testing**, where examinees must obtain a total score above a certain threshold in order to demonstrate a minimum level of competence. Examples of mastery tests include both the knowledge and road portions of tests for receiving a driver's licence, a classroom test for mastery of basic numerical skills for grade 2 students, and a bar examination intended to measure the knowledge and skills every attorney should have before they are granted a license to practice law. The final outcome of mastery testing is often dichotomous, specifically, pass versus fail for, respectively, mastery versus nonmastery.

Content validity is a key concern for mastery tests. It is established through expert opinion over three basic steps: (1) Relevant professionals, such as curriculum specialists or highway safety engineers, are surveyed about the ranges of critical skills or knowledge required for mastery in a particular area. (2) Test items are constructed that should correspond to those competencies, and then test content is reviewed by experts for its representativeness. Given a final draft of the text, the last step is to (c) establish minimum scores required to support the hypothesis of mastery. Haynes et al. (1995) described methods for collecting and summarizing expert opinion about test content.

2. **Internal test structure evidence** concerns the requirement that observed patterns of response consistency or covariances among test scores, such as from battery tests with multiple subtests, should match hypotheses about what test scores should measure. Results of reliability analyses are relevant for inferences about the cohesiveness or consistency of test content. For example, the observation that values of internal consistency reliability coefficients, such as alpha, are reasonably high support—but do not prove for all the reasons discussed earlier in this primer—that test items were sampled from a common domain. Likewise, the hypothesis that scores from alternate forms of a test measure a common domain is supported by relatively high values of alternate form reliability coefficients.

There is large amount of research literature that dates to the origins of factor analysis in the early 1900s about analyses of covariances among multiple sets of scores generated by the same battery test or over different tests. For instance, some IQ tests feature 10 to 15 or so subtests that are administered to examinees over sessions that can last 1–2 hours. It might be expected that, for example, subtest intercorrelations should be positive because all their scores reflect general cognitive ability. If it is also observed that scores from certain subtests covary higher with each other than with other subtests in the test battery, the hypothesis that intelligence does not correspond to a single general factor (i.e., unidimensionality) may also be supported. The SEM technique of CFA is widely used to analyze covariances from multiple measures in order to test hypotheses about the structure or organization of intelligence, personality, or attitudes, among many other possible domains.

3. **Covariance evidence** directly corresponds to the idea of criterion-related validity: Test scores should covary with external variables in anticipated ways, especially when an external variable is seen as a gold-standard criterion, such as an outcome measure, with which the test should substantially covary. For instance, scores from employment screening tests administered to applicants should predict relevant aspects of successful job performance. Another example is scores from an admissions test for an advanced educational or training program. External validity is often evaluated by calculating regression coefficients that estimate the direction and magnitude of associations between test scores and external variables. Unreliability in either test or criterion scores can greatly distort values of regression coefficients in either direction, that is, observed values

are systematically too low or too high due to biasing effects of measurement. Thus, criterion scores should be precise, too.

4. **Response process evidence** concerns more qualitative than strictly quantitative evidence. This is because it concerns hypotheses about the cognitive or mental processes involved in participants' responses to test content; that is, how participants reason or go about working through the steps of a problem, among other possibilities. For example, in **protocol analysis**, examinees describe how they approach, interpret, or analyze the problem while completing test items. They may be encouraged to think out loud about their reasoning, and text of their musings is later analyzed for clues about key cognitive processes or steps. In computer interface usability research, users might speak about their experiences or frustrations while using a particular computer tool (Li et al., 2012). Eye tracking hardware can also be used to objectively record where examinees are looking and how they scan test stimuli. Urbina (2014) described the qualitative analysis of whether examiners score a test in ways that are consistent with target scoring rubrics stated in test manuals.

Described next are two examples of methods in **modern measurement (test) theory**, including generalizability theory, which extends classical reliability methods, and item response theory, which extends classical methods for estimating psychometrics of individual test items or total scores.

GENERALIZABILITY THEORY AND METHODS

Classical methods for reliability analysis are limited to the study of just two occasions, raters, or alternative forms for estimation of, respectively, test–retest, interrater, or alternate-forms reliability. Also, each of type of corresponding measurement error, such as time, content, or rater sampling error, may be estimated in separate studies. But under **generalizability theory**, or **G-theory**, it is possible to simultaneously estimate different sources of measurement error over two or more times, raters, or forms (item sets), among other possibilities (settings, levels of examiner training, etc.) (Cronbach et al., 1963). Each source of measurement error is called a **facet**, and multiple facets can be studied together in the same generalizability study, **G-study**. It is also possible to estimate **measurement error interaction effects**, or conditional effects on score reliabilities that involve two or more facets. For example, test–retest score reliabilities could be appreciably lower for certain alternate versions of a test than for other versions. Such joint effects of facets can create yet even more imprecision than either source alone.

The basic mathematical model for G-theory is the general linear model of ANOVA. Thus, it may be possible to analyze data from a G-study using a standard ANOVA procedure in a computer program for general statistical analysis. Results from a G-study can be represented by a generalizability coefficient, or **G-coefficient**, which estimates how facets combine to affect score reliability and is analogous to a reliability coefficient in classical test theory. Its basis is that of an **intraclass correlation coefficient** (ICC), which can be calculated from an ANOVA source table and serves as an expected lower bound to the target coefficient of generalizability (Cronbach et al., 1963). The same results can also inform a **D-study**, where the effects of alternative measurement plans are estimated. For instance, the effects on score precision of using different numbers of raters (e.g., 2 vs. ≥3) or tests of different lengths (e.g., 20 vs. 40 items) can be approximated in a D-study. See Thompson (2003, chaps. 2–3) for an extended introduction to generalizability theory, and Briesch et al. (2014) offer practical suggestions for applying G-theory in educational settings.

ITEM RESPONSE THEORY AND ITEM CHARACTERISTIC CURVES

For two reasons, it is worthwhile to know about **item response theory** (IRT), also known as **latent trait theory**. First, techniques in IRT permit more sophisticated estimation of item psychometrics than is possible in classical measurement theory. Methods in IRT can be used to equate scores from one test to another, evaluate the extent of item bias over different populations, and construct individualized tests for examinees of different ability levels, or **tailored testing**, among other possibilities. Second, it is an alternative to CFA for analyzing ordinal data. In the past, researchers who analyzed IRT models used specialized software, but now some SEM computer programs such as LISREL and Mplus can analyze at least basic kinds of IRT models. How to analyze ordinal data in CFA is considered later in the book, but part of the logic for doing so is related to that of IRT.

The body of IRT consists of mathematical models that relate responses on individual items to a continuous latent variable θ. Assume for this discussion that items are dichotomously scored (0 = incorrect, 1 = correct) and that θ is an ability dimension with a normal deviate (*z*) metric. Presented in Figure 4.4 is an **item characteristic curve** (ICC), or a sigmoid function that relates θ to the estimated probability of a correct answer. This ICC depicts a **two-parameter IRT model**, where the parameters are item difficulty and item discrimination. Difficulty is the level of ability that corresponds to a 50% chance of getting the item correct, and discrimination is the slope of the tangent line to the ICC at that point. In the figure, difficulty is θ = 0 (i.e., the mean) because this level of ability predicts that 50% of examinees will pass the item, and discrimination is the slope of the tangent line at this point. The steeper the slope, the more discriminating the item, and the stronger its relation with θ. **Three-parameter IRT models** also include a guessing parameter, and it indicates the probability that an examinee of low ability would correctly guess the answer. A **Rasch model** has a single parameter, item difficulty. Uniform discrimination for all items implies a constant construct, one that can be measured in the same way for all examinees regardless of ability level. In this way, evaluation of Rasch models can be viewed as more confirmatory than fitting more complex IRT models to the data.

Figure P.1 might look familiar. This is because the shape of an ICC and the sigmoid functions analyzed in logistic regression and probit regression for dichotomous variables are similar (see Figure R.4). Shared among all these techniques is the analysis of a continuous latent variable that underlies responses to dichotomous observed variables. Parameter estimates in IRT can be scaled in either logistic units or probit units, and we will see later in the book that estimates in CFA can be mathematically transformed to estimates of the type generated in IRT. Baylor et al. (2011) gives a clear introduction to IRT.

SUMMARY

In written reports, researchers should provide information about the psychometrics of their scores. Analysis of scores with poor reliability or validity can jeopardize the results. Because reliability is not a property of tests, best practice is to estimate the reliability of scores analyzed in a particular sample and report those results in written summaries. Values of reliability coefficients derived in other samples can be reported, too, but also directly compare your sample with those other samples. Score reliability is a requirement for validity, but does not guarantee it. Validity concerns the accuracy

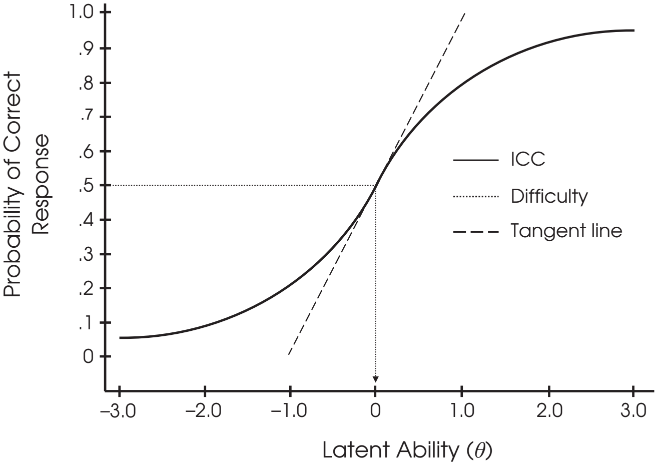

FIGURE P.1. Item characteristic curve (ICC) for the predicted probability of a correct response for a dichotomously scored item in a two-parameter item response theory model. Item difficulty is θ = 0, and item discrimination is the slope of the tangent line at θ = 0.

of interpretations of test scores in a particular setting, including the intended use of the test. Test scores may be valid for one purpose or setting, but not in another, so validity is also not a fixed characteristic of tests. Analysis of scores with poor reliability or validity can jeopardize the results. There are ways in SEM to take account direct account of less-than-perfect (i.e., < 1.0) score reliabilities for observed variables. The capability to do so is a major advantage of SEM over statistical methods for observed variables, such as multiple regression and ANOVA.

LEARN MORE

Urbina (2014) offers a concise introduction to psychometrics at the undergraduate level, and Furr (2022) does so at the graduate level. Thompson (2003) deals with both classical reliability and generalizability.

Furr, R. (2022). *Psychometrics: An introduction* (4th ed.). Sage.

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues.* Sage.

Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). Wiley.

EXERCISES

1. Comment on this statement: A test–retest reliability coefficient of 1.0 means that all cases attained the same score on each of two different occasions.

2. Presented next are scores for 10 cases (*S*) on four variables:

	*i*1	*i*2	*i*3	*i*4
S_1	16	48	100	45
S_2	14	47	92	30
S_3	16	45	88	38
S_4	12	45	95	32
S_5	18	46	98	41
S_6	18	46	101	39
S_7	13	47	97	38
S_8	16	48	98	44
S_9	18	49	110	46
S_{10}	22	49	105	45

Calculate the split-half reliability coefficients for an odd–even split and also for a first half–half, second–half split. Comment on the results.

3. Calculate α for the dataset in Exercise 1. Comment on the results.

4. Calculate α, given the summary statistics for three variables listed next:

$$SD_1 = 2.50,\ SD_2 = 5.00,\ SD_3 = 4.50$$
$$s_1^2 = 6.25,\ s_2^2 = 25.00,\ s_3^2 = 20.25$$
$$r_{12} = .40,\ r_{13} = .60,\ r_{23} = .50$$

5. Prove that $\alpha_S = .95$, given $N_i = 2$ and $r_{12} = .905$. (Because there are only two items, $r_{12} = \overline{r_{ij}}$ for this exercise.)

6. You are developing a group-administered test of math skills for grade 4 students. There is good evidence for score reliability. Identify relevant types of evidence for validity.

ANSWERS

1. Test–retest r_{XX} = 1.0 says only that the absolute positions of scores are perfectly maintained over the two occasions. One possibility is that every case attained the same score on both occasions, but any other pattern that fully preserves absolute differences will also generate r_{XX} = 1.0. For example, if all cases improved by 5 points at the second occasion, no case attained the same score at both times, but r_{XX} = 1.0.

2. Listed next are total scores on test halves:

	Odd	Even	1st ½	2nd ½
S_1	116	93	64	145
S_2	106	77	61	122
S_3	104	83	61	126
S_4	107	77	57	127
S_5	116	87	64	139
S_6	119	85	64	140
S_7	110	85	60	135
S_8	114	92	64	142
S_9	128	95	67	156
S_{10}	127	94	71	150

Odd–even: r_{hh} = .813; r_{11} = 2(.813)/(1 + .813) = .897

First half–second half: r_{hh} = .820; r_{11} = 2(.820)/(1 + .820) = .901

As expected, the two split-half coefficients are not equal because they are based on different splits of the items. Overall, roughly 10% of observed variation in score is due to content sampling error over the halves and the method of splitting the items.

3. Values of summary statistics for the raw data in Exercise 1 are reported next. Results for variances and standard deviations are listed for variables *i*1–*i*4, respectively:

Variances:	8.456, 2.222, 38.933, 30.622
Standard deviations:	2.908, 1.491, 6.240, 5.534

Correlations: r_{12} = .513, r_{13} = .599, r_{14} = .695,
r_{23} = .753, r_{24} = .687, r_{34} = .711

Covariances: c_{12} = 2.222, c_{13} = 10.867, c_{14} = 11.178
c_{23} = 7.000, c_{24} = 5.667, c_{34} = 24.533

$\overline{s}_i^2$ = 20.058; $\overline{c}_{ij}^2$ = 10.244

$$\alpha = 4(10.244)/[20.058 + (4 - 1)10.244] = .807$$

Thus, about 1 – .807, or 19.3% of the observed variance is due to the combination of inconsistent responding at the item level and test length. As expected, values of α and split-half reliability coefficients for the same data are not equal.

4. Results based on the summary statistics for N_i = 3 items for this question are listed next:

$\overline{s}_i^2$ = (6.25 + 25.00 + 20.25)/3 = 17.167

Covariances: c_{12} = 2.50(5.00).40 = 5.000
c_{13} = 2.50(4.50).60 = 6.750
c_{23} = 5.00(4.50).50 = 11.250
$\overline{c}_{ij}^2$ = (5.000 + 6.750 + 11.250)/3
= 7.667

$$\alpha = 3(7.667)/[17.167 + (3 - 1)7.667] = .707$$

5. For N_i = 2 and r_{12} = .905,

$$\alpha_S = 2(.905)/[1 + (2 - 1).905] = .950$$

6. Evaluate content validity; for example, ask education experts about whether item content is representative. Scores on the new test should covary with those from other arithmetic tests, that is, evaluate convergent validity. The new test should also predict later math skills, such as in grade 6; that is, assess predictive validity. Scores on the new tests should not be correlated too highly (e.g., > .90) with scores on, say, reading comprehension tests; that is, evaluate discriminant validity.

REFERENCES

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

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force report. *American Psychologist, 73*(1), 3–25.

Baylor, C., Hula, W., Donovan, N. J., Doyle, P. J., Kendall, D., & Yorkston, K. (2011). An introduction to item response theory and Rasch models for speech–language pathologists. *American Journal of Speech–Language Pathology, 20*, 243–259.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52*(1), 13–35.

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (2019). *The twenty-first Mental Measurements Yearbook*. Buros Institute of Mental Measurements.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J., Rajaratnam, N., & Gelser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement, 61*(4), 517–531.

Furr, R. (2022). *Psychometrics: An introduction* (4th ed.). Sage.

Gulliksen, H. (1950). The reliability of speeded tests. *Psychometrika, 15*(3), 259–269.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*(3), 177–189.

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 1–27). Elsevier.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kline, R. B. (2020). *Becoming a behavioral science researcher: A guide to producing research that matters* (2nd ed.). Guilford Press.

Li, A. C., Kannry, J. L., Kushniruk, A., Chrimes, D., McGinn, T. G., Edonyabo, D., & Mann, D. M. (2012). Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. *International Journal of Medical Informatics, 81*(11), 761–772.

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4*(2), 192–211.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433.

Messick, S. (1995). Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Mulsant, B. H., Kastango, K. B., Rosen, J., Stone, R. A., Mazumdar, S., & Pollock, B. G. (2002). Interrater reliability in clinical trials of depressive disorders. *American Journal of Psychiatry, 159*(9), 1598–1600.

Osborne, J. W. (2013). *Best practices in data screening.* Sage.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99–103.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55.

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues.* Sage.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*(2), 174–195.

Tierney, R. D. (2016). Fairness in educational assessment. In M.A. Peters (Ed.), *Encyclopedia of educational philosophy and theory* (pp. 793–798). Springer.

Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). Wiley.

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization. *Measurement and Evaluation in Counseling and Development, 44*(3), 159–168.

Williams, M. N., Grajales, C. A. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research, and Evaluation, 18*, Article 11.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficient alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*(1), 21–29.