CHAPTER 5

Classical Test Theory

Assumptions, Equations, Limitations, and Item Analyses

Classical test theory (CTT) has been the foundation for measurement theory for over 80 years. The conceptual foundations, assumptions, and extensions of the basic premises of CTT have allowed for the development of some excellent psychometrically sound scales. This chapter outlines the basic concepts of CTT as well as highlights its strengths and limitations.

Because total test scores are most frequently used to make decisions or relate to other variables of interest, sifting through item-level statistics may seem tedious and boring. However, the total score is only as good as the sum of its parts, and that means its items. Several analyses are available to assess item characteristics. The approaches discussed in this chapter have stemmed from CTT.

Classical Test Theory

Classical test *theory* is a bit of a misnomer; there are actually several types of CTTs. The foundation for them all rests on aspects of a total test score made up of multiple items. Most classical approaches assume that the raw score (X) obtained by any one individual is made up of a true component (T) and a random error (E) component:

$$X = T + E. \tag{5–1}$$

The true score of a person can be found by taking the mean score that the person would get on the same test if they had an infinite number of testing sessions.

Because it is not possible to obtain an infinite number of test scores, *T* is a hypothetical, yet central, aspect of CTTs.

Domain sampling theory assumes that the items that have been selected for any one test are just a sample of items from an infinite domain of potential items. Domain sampling is the most common CTT used for practical purposes. The parallel test theory assumes that two or more tests with different domains sampled (i.e., each is made up of different but parallel items) will give similar true scores but have different error scores.

Regardless of the theory used, classical approaches to test theory (and subsequently test assessment) give rise to a number of assumptions and rules. In addition, the overriding concern of CTTs is to cope effectively with the random error portion (*E*) of the raw score. The less random error in the measure, the more the raw score reflects the true score. Thus, tests that have been developed and improved over the years have adhered to one or another of the classical theory approaches. By and large, these tests are well developed and quite worthy of the time and effort that have gone into them. There are, however, some drawbacks to CTTs and these will be outlined in this chapter as well.

Theory of True and Error Scores: Description and Assumptions

The theory of true and error scores has several assumptions; the first, as was already noted, is that the raw score (*X*) is made up of a true score (*T*) plus random error (*E*). Let's say I was to administer the Team Player Inventory (TPI; Kline, 1999) to myself every day for two years. Sometimes my score would be higher and sometimes lower. The average of my raw scores ($\overline{X}$), however, would be the best estimate of my true score (*T*).

It is also expected that the random errors around my true score would be normally distributed. That is, sometimes when I took the TPI my scores would be higher (maybe I was in a good mood or had just completed a fantastic team project that day), and sometimes when I took the TPI my scores would be lower (maybe I was tired, distracted, or had just completed a team project that was a flop). Because the random errors are normally distributed, the expected value of the error (i.e., the mean of the distribution of errors over an infinite number of trials) is 0. In addition, those random errors are uncorrelated with each other; that is, there is no systematic pattern to why my scores would fluctuate from time to time. Finally, those random errors are also uncorrelated to the true score, *T,* in that is there is no systematic relationship between a true score (*T*) and whether or not that person will have positive or negative errors. All of these assumptions about the random errors form the foundations of CTT.

The standard deviation of the distribution of random errors around the true score is called the *standard error of measurement.* The lower it is, the more tightly packed around the true score the random errors will be. Therefore, one index of the degree of usefulness of the TPI will be its standard error of measurement. Now, you

may be thinking, why on earth would anyone want to take the TPI every day for two years? Good question. This is to simulate the notion of taking a test an infinite number of times (50 times, even, would seem pretty infinite!).

An extremely important shift in this approach came when psychometricians demonstrated that the theory of true and error scores developed over *multiple samplings* of the *same person* (i.e., taking the TPI myself 1,000 times) holds over to a *single administration* of an instrument over *multiple persons* (i.e., administering the TPI to a group of 1,000 different people once). The mathematical proofs for this will not be reviewed but can be found in some psychometrics texts (e.g., Allen & Yen, 1979). This new approach speeds things up dramatically, because through the proofs it is possible to collect data once (single administration) on a sample of individuals (multiple persons). The same CTT and assumptions of the true and error scores can now be applied to this sample of TPI scores.

In the latter scenario, of interest is the variance of the raw scores, true scores, and random error across the sample. So instead of taking the TPI for two years to get an estimate of the standard error for one person (e.g., me), I can give it once to 1,000 people and get the same standard error of measurement that will generalize to the population. The equation for this process is as follows:

$$\text{VAR}(X) = \text{VAR}(T) + \text{VAR}(E). \tag{5–2}$$

Given this, it can be shown that the variance of the observed scores VAR(X) that is due to true score variance VAR(T) provides the reliability index of the test (Equation 5–3).

$$\text{VAR}(T)/\text{VAR}(X) = R. \tag{5–3}$$

When the variance of true scores is high relative to the variance of the observed scores, the reliability (R) of the measure will be high (e.g., 50/60 = 0.83), whereas if the variance of true scores is low relative to the variance of the observed scores, the reliability (R) of the measure will be low (e.g., 20/60 = 0.33). Reliability values range from 0.00 to 1.00. Rearranging the terms from the above equations, it can be shown that

$$R = 1 - [\text{VAR}(E)/\text{VAR}(X)]. \tag{5–4}$$

That is, the reliability is equal to 1 – the ratio of random error variance to total score variance. Further, there are analyses that allow for an estimation of R (reliability), and, of course, calculating the observed variance of a set of scores is a straightforward process. Because R and VAR(X) can be calculated, VAR(T) can be solved for with the following equation:

$$\text{VAR}(T) = \text{VAR}(X) \times R. \tag{5–5}$$

It is worth reiterating here that CTTs are largely interested in modeling the random error component of a raw score. Some error is *not* random; it is systematic.

Much time and effort has been spent to identify and deal with systematic error in the context of test validity. However, it remains largely undetermined in CTT. As such, systematic errors (such as changes in scores over time due to learning, growth, training, or aging) are not handled well in CTT.

Ramifications and Limitations of Classical Test Theory Assumptions

Embretson and Reise (2000) review the ramifications (or "rules," as they call them) of CTTs. The first is that the standard error of measurement of a test is consistent across an entire population. That is, the standard error does not differ from person to person but is instead generated by large numbers of individuals taking the test, and it is subsequently generalized to the population of potential test takers. In addition, regardless of the raw test score (high, medium, or low), the standard error for each score is the same.

Another ramification is that as tests become longer, they become increasingly reliable. Recall that in domain sampling, the sample of test items that makes up a single test comes from an infinite population of items. Also recall that larger numbers of subjects make the statistics generated by that sample more representative of the population of people than would a smaller sample. These statistics are also more stable than those based on a small sample. The same logic holds in CTT. Larger numbers of items better sample the universe of items and statistics generated by them (such as mean test scores) are more stable if they are based on more items.

Multiple forms of a test (e.g., Form A and Form B) are considered to be parallel only after much effort has been expended to demonstrate their equality (Gulliksen, 1950). Not only do the means have to be equal but also the variances and reliabilities, as well as the relationships of the test scores to other variables. Another ramification is that the important statistics about test items (e.g., their difficulty) depend on the sample of respondents being representative of the population. As noted earlier, the interpretation of a test score is meaningless without the context of normative information. The same holds true in CTT, where statistics generated from the sample can only be confidently generalized to the population from which the sample was drawn.

True scores in the population are assumed to be (a) measured at the interval level and (b) normally distributed. When these assumptions are not met, test developers convert scores, combine scales, and do a variety of other things to the data to ensure that this assumption is met. In CTT, if item responses are changed (e.g., a test that had a 4-point Likert-type rating scale for responses now uses a 10-point Likert-type rating scale for responses), then the properties of the test also change. Therefore, it is unwise to change the scales from their original format because the properties of the new instrument are not known.

The issues around problems with difference and change scores that were discussed in an earlier chapter have their roots in CTT. The problem is that the changes

in scores from time one to time two are not likely to be of the same magnitude at all initial levels of scores at time one. For example, suppose at time one, a test of math skills is given (a math pretest) to a group of 100 students. They then have four weeks of math training, after which they are given a posttest. It is likely that there will be more gain for those students who were lower on the test at time one than for those who were higher at time one.

In addition, if item responses are dichotomous, CTT suggests that they should not be subjected to factor analysis. This poses problems in establishing the validity for many tests of cognitive ability, where answers are coded as correct or incorrect.

Finally, once the item stems are created and subjected to content analysis by the experts, they often disappear from the analytical process. Individuals may claim that a particular item stem is biased or unclear, but no statistical procedures allow for comparisons of the item content, or stimulus, in CTT.

Item Analysis Within Classical Test Theory: Approaches, Statistical Analyses, and Interpretation

The next part of this chapter is devoted to the assessment of test items. The approaches presented here have been developed within the theoretical framework of CTT. At the outset, it will be assumed that a test is composed of a number of items and has been administered to a sample of respondents. Once the respondents have completed the test, the analyses can begin. There are several pieces of information that can be used to determine if an item is useful and/or how it performs in relation to the other items on the test.

Descriptive Statistics. Whenever a data set is examined, descriptive statistics come first, and the most common of these are the mean and variance. The same is true for test items. The means and standard deviations of items can provide clues about which items will be useful and which ones will not. For example, if the variance of an item is low, this means that there is little variability on the item and it may not be useful. If the mean response to an item is 4.5 on a 5-point scale, then the item is negatively skewed and may not provide the kind of information needed. Thus, while it is not common to examine item-level descriptive statistics in most research applications, in creating and validating tests it is a crucial first step. Generally, the higher the variability of the item and the more the mean of the item is at the center point of the distribution, the better the item will perform.

Means and variances for items scored on a continuum (such as a five-point Likert-type scale) are calculated simply the way other means and variances are calculated. For dichotomous items, they can be calculated in the same way, but there are derivations that provide much simpler formulae.

The mean of a dichotomous item is equal to the proportion of individuals who endorsed/passed the item (denoted p). The variance of a dichotomous item is calculated by multiplying $p \times q$ (where q is the proportion of individuals who failed, or did not endorse, the item). The standard deviation, then, of dichotomous items

is simply the square root of $p \times q$. So, for example, if 500 individuals respond to a yes/no item and 200 respond "yes," then the p value for that item is 200/500, or 0.40. The q is 0.60 (1.0 – 0.40 = 0.60). The variance of the item is 0.24 (0.40 × 0.60 = 0.24) and the standard deviation is the square root of 0.24, or 0.49.

Difficulty Level. As noted above, the proportion of individuals who endorse or pass a dichotomous item is termed its p value. This might be somewhat confusing because p has also been used to denote the probability level for a calculated statistic given a particular sample size. To keep them separated, it will be important to keep in mind the context in which p is being used. For this section of the book on item difficulty, p will mean the proportion of individual respondents in a sample that pass/endorse an item.

It is intuitive to grasp that on an achievement test, one can pass an item. It is also the case that many tests of individual differences ask the respondent to agree or disagree with a statement. For example, I might want to assess extroversion by asking the respondent a series of questions that can be answered with a yes (equivalent to a pass on an achievement test) or no (equivalent to a fail on an achievement test). An example of such an item would be, "I enjoy being in social situations where I do not know anyone." The respondent then responds yes or no to each of these types of items. A total score for extroversion is obtained by summing the number of yes responses.

While p is useful as a descriptive statistic, it is also called the item's difficulty level in CTT. Items with high p values are easy items and those with low p values are difficult items. This carries very useful information for designing tests of ability or achievement. When items of varying p values are added up across all items, the total (also called composite) score for any individual will be based on how many items she or he endorsed, or passed.

So what is the optimal p level for a series of items? Items that have p levels of 1.00 or 0.00 are useless because they do not differentiate between individuals. That is, if everyone passes an item, it acts the same as does adding a constant of 1 to everyone's total score. If everyone fails an item, then a constant of 0 is added to everyone's score. The time taken to write the item, produce it, respond to it, and score it is wasted.

Items with p values of 0.50—that is, 50% of the group passes the item—provide the highest levels of differentiation between individuals in a group. For example, if there are 100 individuals taking a test and an item has a p value of 0.50, then there will be 50×50 (2,500) differentiations made by that item, as each person who passed is differentiated from each person who failed the item. An item with a p value of 0.20 will make 20×80 (1,600) differentiations among the 100 test takers. Thus, the closer the p value is to 0.50, the more useful the item is at differentiating among test takers.

The one caveat about the p value of 0.50 being the best occurs when items are highly intercorrelated. If this is the case, then the same 50% of respondents will pass all of the items and one item, rather than the entire test, would have sufficed to differentiate the test takers into two groups. For example, assume I have a class of 20 people and give them a 10-item test comprised of very homogeneous items. Further

assume that the *p* value for all 10 items is 0.50. The same 50% of the 20 students would pass all of the items as would pass only one item. Therefore, this test of 10 items is not any better than a test of one item at differentiating the top and bottom 50% of the class. It is because of this characteristic that test designers usually attempt to create items of varying difficulty with an average *p* value across the items of 0.50 (Ghiselli, Campbell, & Zedek, 1981).

Some tests are designed deliberately to get progressively more difficult. That is, easy questions are placed at the beginning of a test and the items become more and more difficult. The individual taking the test completes as many items as possible. These adaptive tests are used often in individual testing situations such as in the Wechsler Adult Intelligence Scale (Wechsler, 1981) and in settings where it can be of value to assess an individual quickly by determining where that person's cutoff point is for passing items. Rather than giving the person the entire test with items of varying levels of difficulty interspersed throughout, whether or not the person passes an item determines the difficulty of the next item presented.

Sometimes instructors deliberately put a few easy items at the beginning of a test to get students relaxed and confident so that they continue on and do as well as possible. Most of us have had the negative experience of being daunted by the first question on a test and the lowered motivation and heightened anxiety this can bring. Thus, instructors should be quite conscious of the difficulty level of items presented early on in a testing situation.

Discrimination Index. Using the *p* values (difficulty indices), discrimination indices (*D*) can be calculated for each dichotomous item. The higher the *D,* the more the item discriminates. Items with *p* levels in the midrange usually have the best *D* values and, as will be demonstrated shortly, the opportunity for *D* to be highest occurs when the *p* level for the item is at 0.50.

The extreme group method is used to calculate *D.* There are three simple steps to calculating *D.* First, those who have the highest and lowest overall test scores are grouped into upper and lower groups. The upper group is made up of the 25%–33% who are the best performers (have the highest overall test scores), and the lower group is made up of the bottom 25%–33% who are the poorest performers (have the lowest overall test scores). The most appropriate percentage to use in creating these extreme groups is to use the top and bottom 27% of the distribution, as this is the critical ratio that separates the tail from the mean of the standard normal distribution of response error (Cureton, 1957).

Step two is to examine each item and determine the *p* levels for the upper and lower groups, respectively. Step three is to subtract the *p* levels of the two groups; this provides the *D.* Table 5.1 shows an example for a set of four items. Assume that these data are based on 500 individuals taking a test that is 50 items in length. The highest scoring 135 individuals (500×0.27) for the entire test and lowest scoring 135 individuals for the entire test now make up our upper and lower extreme groups. For Item 1, the upper group has a *p* level of 0.80 and the lower group has a *p* level of 0.30. The *D,* then, is $0.80 - 0.30 = 0.50$. For Item 2, the *D* is 0.80; for Item 3, it is 0.05; and for Item 4, it is –0.60.

Table 5.1 Example of Item Discrimination Indices

Item	*p Level for Upper Group*	*p Level for Lower Group*	*D*
1	0.80	0.20	0.60
2	0.90	0.10	0.80
3	0.60	0.55	0.05
4	0.10	0.70	–0.60

Items 1 and 2 have reasonable discrimination indices. The values indicate that those who had the highest test scores were more likely to pass the items than individuals with low overall scores. Item 3 is very poor at discriminating; although 60% of those in the upper group passed the item, almost as many (55%) in the lower group passed the item. Item 4 is interesting—it has a negative *D* value. In tests of achievement or ability, this would indicate a poor item in that those who scored most highly on the test overall were not likely to pass the item, whereas those with low overall scores were likely to pass the item. However, in assessment tools of personality, interests, or attitudes, this negative *D* is not problematic. In these types of tests, it is often of interest to differentiate between types or groups, and items with high *D* values (positive or negative) will help in differentiating those groups.

Using p Levels to Plot Item Curves. A technique of item analysis that foreshadowed modern test theory was developed in 1960. The Danish researcher Rasch plotted total test scores against pass rates for items on cognitive tests. These curves summarized how an individual at an overall performance level on the test did on any single item. Item curves using *p* levels provided more fine-grained information about the item than just the *p* level overall or the discrimination index did.

Figure 5.1 shows examples of item curves for four separate items. Assume that the entire test was 50 items in length and that 200 students took the test. Then the 200 students were separated into percentiles with cutoffs placed at the 10th, 20th, and so forth to the 99th percentile performance for the entire test. Each of the percentiles is plotted against the *p* level associated with that item for that percentile.

Note that for Item 1, as the performance of the students on the test increased (i.e., they are in the higher percentile groups), the performance on the item increased. However, at the 50th percentile, the increased *p* levels slowed. This indicates that up to the 50th percentile level, as overall test performance increased, the pass rate increased in a monotonic linear fashion. After the 50th percentile, as test performance increased, so did the pass rate for the item. However, the curve is not as steep, and thus the item did not discriminate between individuals as well at the upper end of the overall test distribution as it did for those at the lower end.

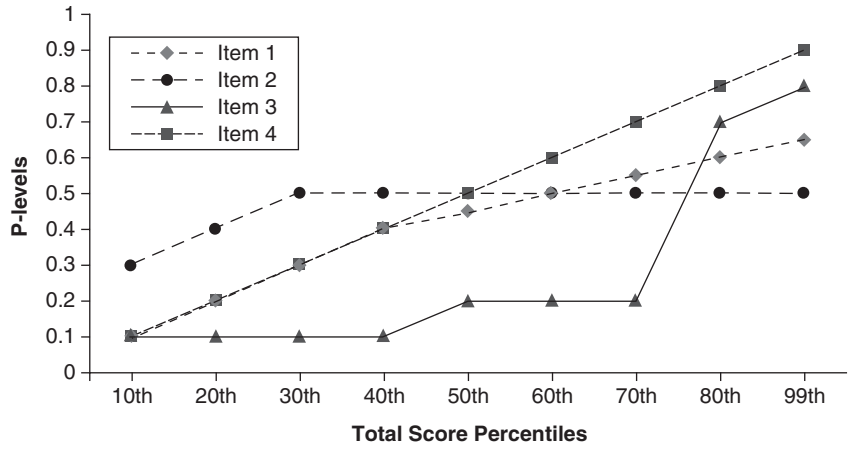

Figure 5.1 Item Curves Based on *p* Levels

Item 2 starts off with more difficulty because the *p* level for the lowest group is 0.30. The slope moves up over the next two percentile levels but after that, the line is flat. This indicates a relatively poorly discriminating item at the lower end of the performance continuum and a nondiscriminating item at the mid- to upper-performance levels.

Item 3 shows relatively little difference in *p* level until one gets to the 80th percentile level. At this point, there is a sharp spike, indicating that the item is particularly good at discriminating between the 70th and 80th percentile students. Finally, Item 4 is a straight line. As the overall performance increases (i.e., percentile) the *p* level increases at a steady rate.

Item-to-Total Correlations. Another assessment of items related to its discrimination index is the Pearson product-moment item-to-total correlation coefficient. For dichotomous items, the Pearson point-biserial or Pearson biserial correlation coefficients are available. The underlying question addressed by each coefficient is the same: How do responses to an item relate to the total test score?

For all three statistics, the relationships between how individuals responded to each item are correlated with the *corrected* total score on the test. The correction is made insofar as the total score does not include the response to the item in question. This is an appropriate correction because total scores that have the item in question embedded within them will have a spuriously higher relationship (i.e., correlation) than total scores made up of only the other items in the test. This correction is particularly important when there are only a few test items—say five or six. However, if a test has 100 items, the influence of any one item on the total score is minimal. There is no rule for how many items should be included before the item has little influence, so it is better to be conservative in the estimates and use the corrected score.

Which version of the Pearson is appropriate? Assume there are 10 items in a scale and each is responded to on a seven-point Likert-type scale. Responses to each item are then correlated to the corrected total scores for each test taker. This is the same as having two continuous variables, and the Pearson product-moment correlation is the right one to use. Table 5.2 shows an example of the vector for one item that is responded to on a four-point Likert-type scale (strongly disagree = 1, disagree = 2, agree = 3, and strongly agree = 4) and a vector of the corrected total scores on a 10-item test across 20 participants.

Table 5.2 Two Continuous Variables Used in Calculating Item-to-Total Correlations for Item 1 of a 10-Item Test

Participant	*Four-Point Likert-Type Response*	*Total Score*[a]
1	3	34
2	4	30
3	4	32
4	2	15
5	3	20
6	3	27
7	4	31
8	1	12
9	4	23
10	3	25
11	2	18
12	1	11
13	1	15
14	3	27
15	2	20
16	2	19
17	2	20
18	1	16
19	4	25
20	1	32

a. The total score is corrected so that it does not include the score from Item 1.

One of the hand calculations for the Pearson product-moment correlation coefficient when the variance of the variables is readily at hand is

$$r = [\Sigma XY/n - (\overline{X})(\overline{Y})]/(\sigma_x \times \sigma_y), \tag{5–6}$$

where $\Sigma XY/n$ = the mean of the sum of cross-products of variable X (item score) and variable Y (total score), $\overline{X}$ = the mean of the scores on the X variable, $\overline{Y}$ = the mean of the scores on the Y variable, σ_x = the standard deviation of scores on the X variable, and σ_y = the standard deviation of scores on the Y variable.

Substituting the appropriate values from Table 5.2, we get the following equation:

$$\begin{aligned} r &= [61.65 - (2.5)(22.6)]/(1.15 \times 7.04), \\ &= 5.15/8.10, \\ &= 0.64. \end{aligned}$$

Thus, the item-to-total correlation for this item is 0.64.

If the responses represent a true dichotomy (e.g., yes/no; agree/disagree; pass/fail), then this means there is a vector of 1s and 0s for each of the items and a continuous score for the total score on the test. A true dichotomy is one where the categorization really has only two possible alternatives for a single item (e.g., male/female; married/single; yes/no; pass/fail). In this case, the Pearson point-biserial item-to-total correlation coefficient is the appropriate statistic.

If the responses represent a false dichotomy, then there will still be a vector of 1s and 0s for each of the items and a continuous score for the total score on the test. In this instance, however, the Pearson biserial item-to-total correlation coefficient is the appropriate statistic. A false dichotomy is one where an arbitrary decision has been made to force a continuous variable into a dichotomous one. For example, if someone passes or fails a test, the test taker does so because she or he has made or not made it past a particular cutoff score. That cutoff score is arbitrarily set. Similarly, if scores on a four-point Likert-type scale (strongly disagree, disagree, agree, and strongly agree) are grouped into two categories (strongly agree and agree = 1; disagree and strongly disagree = 0), this is a false dichotomy.

It is important to note that computer programs will *not* recognize the difference between 1s and 0s that represent a true dichotomy and those that represent a false dichotomy. It is up to the researcher to know the difference and specify the correct analysis. If the point-biserial equation is used when the biserial was supposed to be, it will underestimate the true strength of the relationship. This is because the biserial correlation takes into account that underlying the 1s and 0s is a normal distribution of scores. One popular computer program (SPSS) does not calculate biserial correlation coefficients. However, the hand calculations of the point-biserial and biserial correlations are not difficult. An example of how to carry them out is shown next.

Table 5.3 shows a vector of dichotomous item responses and a vector of the corrected total scores on a 10-item test across 20 participants. First, assume that the responses are 1 = yes and 0 = no; thus, we have a true dichotomy and use the point-biserial correlation coefficient.

The hand-calculation formula for the point-biserial correlation coefficient is

$$r_{pbis} = [(\overline{Y}_1 - \overline{Y})/\sigma_y] \times \sqrt{p_x/q_x} \tag{5–7}$$

where $\overline{Y}_1$ = the mean of the total test scores for those whose dichotomous response was 1, $\overline{Y}$ = the mean of the total test scores for the whole sample, σ_y = the standard deviation of all scores on the total test, p_x = the proportion of individuals whose dichotomous response was 1, and q_x = the proportion of individuals whose dichotomous response was 0.

Table 5.3 One Dichotomous and One Continuous Variable Used in Calculating Item-to-Total Correlations for Item 1 of a 10-Item Test

Participant	*Dichotomous Response*	*Total Score*[a]
1	1	9
2	1	8
3	1	7
4	0	5
5	1	6
6	1	4
7	1	7
8	0	2
9	1	5
10	1	8
11	0	3
12	0	2
13	0	4
14	1	5
15	0	1
16	0	3
17	0	2
18	0	4
19	1	9
20	0	2

a. The total score is corrected so that it does not include the score from Item 1.

Substituting the correct values into the equation,

$$\begin{aligned} r_{pbis} &= [(6.8 - 4.8)/2.53] \times \sqrt{0.5/0.5}, \\ &= (2/2.53) \times \sqrt{1}, \\ &= 0.79 \times 1, \\ &= 0.79. \text{ Thus, the item-to-total correlation for this item is 0.79.} \end{aligned}$$

Now assume that the responses to the "dichotomous" item have been converted from the responses in Table 5.2, where a 1 or 2 = 0 and a 3 or 4 = 1. This is the *same* data that was used to calculate the point-biserial, but the data represent a false dichotomy, so the biserial correlation coefficient is needed. The hand-calculation formula for the biserial correlation coefficient is

$$r_{bis} = [(\overline{Y}_1 - \overline{Y})/\sigma_y)] \times (p_x/\text{ordinate}), \tag{5–8}$$

where $\overline{Y}_1$ = the mean of the total test scores for those whose dichotomous response was 1, $\overline{Y}$ = the mean of the total test scores for the whole sample, σ_y = the standard deviation of scores for the whole sample on the total test, p_x = the proportion of individuals whose dichotomous response was 1, and ordinate = the ordinate (y-axis value) of the normal distribution at the z value above which p_x cases fall.

In this case, the p_x is equal to 0.50. The corresponding z value above which 50% of the distribution lies is 0.00. The ordinate value (height of the curve on the y-axis) associated with a z value of 0.00 is 0.3989. So, substituting into the formula,

$$\begin{aligned} r_{bis} &= [(6.8 - 4.8)/2.53] \times (0.50/0.3989), \\ &= 0.79 \times 1.25, \\ &= 0.99. \end{aligned}$$

Thus, the item-to-total correlation for this item is 0.99. Notice that the biserial correlation with the exact same data is much higher (0.99) than the point-biserial value (0.79).

Item assessment interpretation using any of the Pearson correlation coefficients is similar to the usual interpretation of this statistic. For all three versions, the values range from −1.00 to +1.00. Negative values indicate that the item is negatively related to the other items in the test. This is not usually desirable as most tests try to assess a single construct or ability. Items with low correlations indicate that they do not correlate, or "go with," the rest of the items in the data set. In addition, be forewarned that items that have very high or very low p levels have a restriction of range and thus will also have low item-to-total correlations. It is best to have midrange to high item-to-total correlations (say 0.50 and above).

Item-to-Criterion Correlations. Another index of item utility is to examine its relationship with other variables of interest. For example, suppose an item on an interest inventory asks the respondent to indicate yes or no to the following item: "I enjoy studying biology," and this item is administered to a group of high school seniors. Their responses (1 = yes and 0 = no) are correlated with scores on the same item answered by a group of physicians. If the correlation is high, then the item is said to

discriminate between individuals who have similar interests to physicians and those who do not. If the correlation is low, then the item is said not to discriminate between individuals who have similar interests to physicians and those who do not.

As another example, consider an item on an employee selection instrument (e.g., "I always complete my work on time") and correlate it with an aspect of job performance (e.g., "completes jobs assigned in a timely manner"). If the correlation is high, then the item relates well to that aspect of job performance. If the correlation is low, then the item does not relate well to that aspect of job performance.

Inter-Item and Item-to-Criterion Paradox. There is an unusual paradox in scale development and use around the notions of inter-item correlations and correlations between items and external (criterion) variables. That is, if a scale is created that is highly homogeneous, then the items will have high intercorrelations. If a scale is created deliberately to capture heterogeneous constructs so that the items can be related to scores on a multifaceted criterion such as job performance, then the items will have low inter-item correlations but the total score is likely to relate well to the multifaceted criterion. Always be conscious of exactly what the scores on a test are to be used for. When they are used in a manner not consistent with the design of the scale, unusual findings may result.

Here is a concrete example. Suppose you design a test of cognitive reading ability and it is designed to be homogeneous. This test is then used to select graduate students into a program. The criterion measure used is supervisor ratings of overall student performance. When the two are correlated, the relationship between them is small (say 0.15). The problem is not necessarily with the test; the problem is that the criterion is multifaceted and this was not taken into account. Specifically, overall student performance ratings would likely encompass diligence in working on research, performance in statistics as well as verbally loaded courses, teaching ability, ability to get along with others, volunteering for tasks, and so forth. Cognitive reading ability (i.e., your test) would be better related to performance in verbally loaded courses and none of the other tasks.

Differential Item Weighting. Differential item weighting occurs when items are given more or less weight when being combined into a total score. This is in contrast to unit-weighting items, where each item has a weight of 1.0 (i.e., effectively contributing equally to the total score). There are several different options for assigning weights to items (e.g., Ghiselli, Campbell, & Zedek, 1981). The first group of techniques is based on statistical grounds. For example, the reliability of items can be calculated and then the reliabilities can be used to assign different weights to the items. Items with higher reliabilities carry more weight in the total score. Another option would be to use a criterion measure and regress the criterion on the items and use the resulting regression weights as the basis for item weighting. Again, those with higher weights are, in turn, weighted more heavily in generating the total score. Another way to decide on weights is to run a factor analysis and use the factor loadings to assign weights to the items. Finally, item-to-total correlation coefficients can be used to weight the items.

Alternatively, theory or application may drive the decision making, and items that are deemed by some decision rule (e.g., majority or consensus) to be more

important or meaningful are given more weight. For example, if there is a 10-item assessment of instruction for a course, and "organization" and "fairness" are perceived by stakeholders to be more important than "punctuality" or "oral skills," then the items can be weighted accordingly when obtaining a total score on teaching effectiveness.

While much effort goes into discussing and determining differential item weights, Ghiselli, Campbell, and Zedek (1981) are persuasive in arguing that differential item weighting has virtually no effect on the reliability and validity of the overall total scores. Specifically, they say that "empirical evidence indicates that reliability and validity are usually not increased when nominal differential weights are used" (p. 438). The reason for this is that differential weighting has its greatest impact when there (a) is a wide variation in the weighting values, (b) is little intercorrelation between the items, and (c) are only a few items. All three are usually the opposite of what is likely to occur in test development. That is, if the test is developed to assess a single construct, then if the developer has done the job properly, items will be intercorrelated. As a result, the weights assigned to one item over another are likely to be relatively small. In addition, tests are often 15 or more items in length, thus rendering the effects of differential weighting to be minimized. Finally, the correlation between weighted and unit-weighted test scores is almost 1.0. Thus, the take-home message is pretty simple—don't bother to differentially weight items. It is not worth the effort.

Summary

This chapter reviewed the basic tenets of classical test theory and also the types of analyses that can be used to assess items within classical test theory. The chapter covered

a. the assumptions, ramifications, and limitations of classical test theory;
b. interpretations of item descriptive statistics and discrimination indices;
c. plots of item curves using pass rates;
d. correlations between items and the total test score and between items and external criteria; and
e. a brief description and comment on the utility of differential item weighting.

In the next chapter, modern test theory is presented as well as item analyses associated with that theoretical framework.

Problems and Exercises

1. What are the components in the equation $X = T + E$ and what do they mean?
2. What are the components in the equation $R = 1 - [\text{VAR}(E)/\text{VAR}(X)]$ and what do they mean?

3. What are the components in the equation VAR(T) = VAR(X) × R, what do they mean, and why is this equation so important?
4. What is meant by the term *random error* in classical test theory?
5. How is the standard error of measurement of a test interpreted across multiple individuals' test scores in classical test theory?
6. How does the "law of large numbers" regarding test length manifest itself in classical test theory?
7. Calculate the mean, variance, and standard deviation of a test item where there were 200 respondents and 50 passed the item.
8. What is the p level of a dichotomous item and at what value does it make the highest discrimination?
9. How many discriminations in a sample of 500 test takers will an item with a p level of 0.40 make?
10. Calculate the discrimination index for an item with the lowest group having a p level of 0.10 and the highest group having a p level of 0.80.
11. What is "corrected" in an item-to-total correlation coefficient?
12. What is the difference between a point-biserial and a biserial correlation coefficient?
13. Calculate the point-biserial (assuming a true dichotomy) and biserial (assuming a false dichotomy) correlations for an item with the following statistics:

    Mean of the continuous variable using all test taker scores = 10

    Standard deviation of the continuous variable using all test taker scores = 7

    Mean of the continuous variable for those who passed the items = 13

    Proportion of those who passed the item = 0.80

    Proportion of those who failed the item = 0.20

    Ordinate of the z value for 40% = 0.3867
14. If an item on a pretest correlates with overall score of 0.70 (and is significant) in a course, how would one interpret this?