CHAPTER 9

# Assessing Validity Using Content and Criterion Methods

Once the test has been deemed reliable enough to use, it is time to begin the lengthy validation process. Often the analyses to assess an instrument's psychometric soundness will provide evidence for both reliability and validity. In many instances, the two issues are strongly tied. However, from a pedagogical perspective, it is useful to separate those analyses most closely linked with reliability from those most closely linked with validity. Keep in mind, though, that the two psychometric properties are not mutually exclusive.

The term *test validity* is a misnomer. Tests are not valid in and of themselves. The inferences made about a test score are more or less valid. For example, if I have a higher score on a need for achievement scale than does my colleague, then I want to infer that I am higher on the trait of need for achievement. Validity also is concerned with how the scores are used. For example, if I use the need for achievement scores for determining who gets a raise—myself or my colleague—this may not be a valid use of the test scores. This way of conceptualizing validity flies in the face of how validity is typically described, which is to indicate that there are different kinds of validity: face, content, criterion-related, and construct. These have been convenient categories that imply that there are various types of validity and that a test can have some of one category and none of another. This assumption is simply untenable.

The approach to test score validity that will be used in this and the next chapter does not assume there are different kinds of validity. Instead, there are different methods used to assess various aspects of test score validity for certain contexts. This chapter will cover the processes and analyses associated with assessing the wording and content of the items on the test and how test scores are related to other variables. The next chapter will cover the processes and analyses associated with assessing the internal structure of the test items.

## Asking the Test Takers

A group of individuals that will provide valuable input in terms of what the test items ask is the test takers themselves. Test takers are important stakeholders in the testing process. If test takers find the items believable, they are more likely to respond appropriately. In the past, this has been called *face validity.* For example, if a group of employees is told that they will be assessed on their "propensity to want to be involved in teams," and they are administered a questionnaire with items such as, How often do you take office supplies home? the employees will likely have very negative reactions to the testing process.

While this is an extreme example of poor face validity, it is not uncommon for test items to seem irrelevant to test takers. This may be because the test items are poorly designed, the use to which the scores will be put is not of interest to or runs counter to the interest of test takers, or because the test items do not meet the expectations of test takers.

Thus, it is critical that items on a test appear both relevant and clear to the test takers. Pilot testing with the instrument using a small sample of test takers from the population of test takers is a great time-saving exercise. For example, with the Team Player Inventory (TPI), employees who have worked in team environments are the most appropriate ones from whom to solicit feedback as to the utility of the items. Usually after they take the test, extensive debriefing with the test takers as to the clarity of the items and their relevance is carried out. The information gathered assists developers and administrators in knowing whether or not it will even be worthwhile to administer the test to a larger sample.

Test taker comments and suggestions are completely subjective and are made by laypeople. Thus they do not have the aura of being "scientific." However, skipping this step may have dire consequences for using the test. If the test items fail to have adequate face validity, then rewriting the items or selecting a different instrument is warranted.

## Asking the Subject Matter Experts

Another perspective that also falls under the "completely subjective" category is that of Subject Matter Experts (SMEs). SMEs can be grouped into two types for purposes of validation; one is for content and one is for process. Content SMEs were used as part of the initial item-creation exercise. Another group of content SMEs are now needed to determine if, from their perspective, the items capture the construct. As with the test takers, input from SMEs at the front end of test development or administration will save lots of problems later in the validation process.

To get the best feedback from the content SMEs, tell them very clearly what is being measured. By doing so, they will be better informed and able to indicate if the items are deficient (i.e., missing some aspect of the construct that should be included) or are contaminated (i.e., contain items that will solicit information not in the construct).

The process SMEs are the ones who will provide feedback on test administration procedures. They will help to ensure that the questions are clear and unambiguous. They will help decide what would be the most appropriate responding approach to use that will allow for making valid inferences from the test scores. These individuals should also help with administrative issues such as test length, compliance with legal issues, test presentation format, etc. For example, if a test is 400 items long, it may be wonderful in capturing the construct, but respondent fatigue would be at least as important an issue to consider as is item content. If the test is to be used with a special population, such as the elderly, handicapped, children, adolescents, and so forth, then how best to administer the test (one on one, electronically, paper and pencil, etc.) might be a consideration. If the test is speeded, scores on such tests are valid only if the construct being measured has speed as a primary characteristic (e.g., data entry, filing). These are the types of potential problems the process SMEs will help you to avoid.

This type of SME input in the past has been called *content validity.* It is obvious why content validity is subjective in nature and is present in any testing situation to some degree—it does not either exist or not exist. Rather, it is up to test administrators to ensure that they have exercised due diligence in constructing and/or selecting the most appropriate test to use on their particular sample for their particular purpose. SMEs are an invaluable resource and should be used to the fullest extent in test development, design, and administration. Their input should assist in efficiently administering a high-quality test under optimal conditions.

## Assessments Using Correlation and Regression: Criterion-Related Studies

One common way to assess the utility of tests scores is to use them to predict other variables of interest. For example, it might be expected that teams with higher ratings of being team players (using team-aggregated scores on the TPI) would also have higher ratings on their willingness to work together in the future compared to teams with lower TPI scores. It might also be expected that individuals with higher TPI scores would like working on teams and thus have worked on more teams in the past than individuals with lower scores on the TPI. Relationships such as these are called *criterion-related* in that one variable, the predictor (e.g., TPI scores), is being used to predict another variable of interest, the criterion (e.g., number of teams worked on in the past two years). Thus, it follows that this type of validity assessment technique has been called in the past *criterion validity* and the correlations produced are called *validity coefficients* (although this is quite strong language considering much more is associated with validity than just a correlation coefficient!).

Many assessments, such as the type just described, of the validity of test scores use correlation or regression analyses. In Chapter 1, these analyses were described and thus will not be reviewed here again. Table 9.1 shows the TPI items and Table 9.2 the TPI scores for 25 individuals, the number of work teams they had been involved in over the past two years, and their ages.

**Table 9.1** Team Player Items (Kline, 1999)

1. I enjoy working on team/group projects.*
2. Team/group project work easily allows others to not "pull their weight". (R)*
3. Work that is done as a team/group is better than the work done individually.*
4. I do all the work in team/group projects, while others get the credit. (R)
5. Others on the team/group benefit from my input.
6. I do my best work alone rather than in a team/group. (R)*
7. My experiences working in team/group situations have been primarily positive.
8. Working with others in a team/group situation slows my progress. (R)
9. My personal evaluation should be based on my team/group's collective work.
10. Team/group work is overrated in terms of the actual results produced. (R)*
11. Working in a team/group gets me to think more creatively. *
12. Team/groups are used too often, when individual work would be more effective. (R)*
13. A benefit of working in a team/group situation means that I get to meet new people.
14. A problem with working in a team/group situation is that some team members may feel "left out." (R)
15. A benefit of working in a team/group situation is that it gives the members a sense of common purpose.
16. Working in a team/group situation fosters conflict between the teams/groups. (R)
17. My own work is enhanced when I am in a team/group situation.*
18. My experiences working in team/group situations have been primarily negative. (R)*
19. More solutions/ideas are generated when working in a team/group situation than when working alone. *
20. I work harder alone than when I am in a team/group situation. (R)

*Note:* (R) = reverse coded items, * = retained items

If a simple correlation is done between TPI scores and number of work teams involved in over the past two years, the resulting correlation is 0.75. This is significant and indicates that the overlapping variance between TPI scores and work teams involved in is 56%. Table 9.3 shows the output of the SPSS correlation between the two variables.

If the question was phrased in more predictive terminology, the number of work teams involved in over the past two years would be regressed on TPI scores. The results would be the same in terms of the magnitude of the relationship, with a $b = 0.119$ and $\beta = 0.747$. The latter analysis is more directional than the former in that the researcher makes clear which variable is the predictor and which is the criterion. Box 9.1 shows regression output of the data in Table 9.1.

The issue of which analysis to use would be more relevant if there were two or more variables as predictors in the analysis. If this was the case, then regression would be the most appropriate technique to use. For example, assume that both

**Table 9.2** Data for Assessing the Criterion-Related Association Between Team Player Inventory Scores and Number of Work Teams Involved in Over the Past Two Years

| *Case* | *TPI Score* | *Number of Teams* | *Years of Age* |
|---|---|---|---|
| 1 | 10 | 2 | 50 |
| 2 | 12 | 3 | 30 |
| 3 | 14 | 4 | 40 |
| 4 | 15 | 1 | 45 |
| 5 | 16 | 2 | 46 |
| 6 | 18 | 3 | 47 |
| 7 | 22 | 2 | 48 |
| 8 | 24 | 4 | 35 |
| 9 | 25 | 3 | 34 |
| 10 | 27 | 2 | 36 |
| 11 | 28 | 4 | 35 |
| 12 | 29 | 3 | 34 |
| 13 | 31 | 5 | 30 |
| 14 | 33 | 2 | 29 |
| 15 | 35 | 6 | 30 |
| 16 | 37 | 4 | 33 |
| 17 | 40 | 7 | 35 |
| 18 | 41 | 3 | 26 |
| 19 | 42 | 6 | 23 |
| 20 | 43 | 5 | 24 |
| 21 | 44 | 8 | 26 |
| 22 | 46 | 6 | 27 |
| 23 | 47 | 8 | 29 |
| 24 | 47 | 6 | 27 |
| 25 | 49 | 5 | 26 |

TPI scores and age are predictors of the number of teams worked on in the past two years. It is anticipated that older workers will be less likely to have been brought up with teams as typical work units, and, thus, as a conservative scientist, the effects of age need to be controlled for before assessing the predictive utility of the TPI scores.

**Table 9.3** Correlation Between Team Player Inventory Scores and Number of Work Teams Involved in Over the Past Two Years Using SPSS

| | *TPI* | *Number of Work Teams* |
|---|---|---|
| TPI | 1.00 | 0.747 |
| Sig. (two-tailed) | | 0.000 |
| *N* | 25 | 25 |
| Number of Work Teams | 0.747 | 1.0 |
| Sig. (two-tailed) | 0.000 | |
| *N* | 25 | 25 |

*Note:* The correlation between the two variables is 0.747 and is significant at an $\alpha$ of < 0.01. The two variables are positively and significantly correlated.

In this case, a hierarchical regression analysis is run where number of work teams involved in over the past two years is first regressed on age (resulting in an $R^2$ of 0.40, indicating that 40% of the variance of number of teams is accounted for by age). Then number of work teams involved in is regressed on age *and* TPI scores. The resulting $R^2$ is 0.560 and is significant, indicating that age and TPI scores together account for 56% of the variance in number of work teams involved in. Our question of interest, however, is whether the increased percentage (56% − 40% = 16%) by adding TPI scores is significant or not. This change upward in the $R^2$ value is indeed significant, indicating that TPI scores add significantly to predicting number of work teams involved in above and beyond that of age.

Box 9.2 shows the hierarchical regression analyses of the number of work teams involved in regressed on age and TPI scores based on the data in Table 9.2. This hierarchical approach is a more conservative test of the TPI scores in that age is first allowed to account for its shared variance in number of teams worked, and then any variance accounted for by TPI scores in addition to age is assessed for significance.

The examples reviewed above are relatively simple (one or two predictors and one criterion). It is easy to add more predictors, and thus the interpretation becomes more complex. For example, let's say cognitive skill scores, honesty scores, conscientious personality scores, and TPI scores are used to predict job performance ratings by supervisors (see Table 9.4).

In this case, job performance of 50 employees (using supervisor ratings of job performance) is regressed on all four predictor variables simultaneously (termed a simultaneous or direct solution). (See Box 9.3 for a detailed description of the analyses and output associated with this example.)

The overall $R^2$ value in this analysis is 0.641, indicating that all four variables together account for 64% variance in job performance. However, which of the four predictors add incremental information to knowledge of job performance above and beyond the other three variables is also of interest. In this case, the *b* values and their significance as well as the $\beta$ values are examined to assess the relative contribution of each variable in predicting job performance. In this particular data set, the variables that account for significantly unique variance above and beyond the others are cognitive skills and TPI scores.

**Box 9.1** Bivariate Regression of Number of Work Teams Involved in Over the Past Two Years Regressed on Team Player Inventory Scores Using SPSS

| *Model Summary* | | | | |
|---|---|---|---|---|
| *Model* | *R* | *R-Square* | *Adjusted R-Square* | *Std. Error of the Estimate* |
| 1 | 0.747 | 0.558 | 0.538 | 1.33996 |

| *ANOVA* | | | | | |
|---|---|---|---|---|---|
| *Model* | *Sum of Squares* | *df* | *Mean Square* | *F* | *Sig.* |
| Regression | 52.064 | 1 | 52.064 | 28.997 | 0.000 |
| Residual | 41.296 | 23 | 1.795 | | |
| Total | 93.360 | 24 | | | |

| *Coefficients* | | | | | |
|---|---|---|---|---|---|
| | *Unstandardized Coefficients* | | *Standardized Coefficients* | *t* | *Sig.* |
| *Model* | *B* | *Std. Error* | *Beta* | | |
| Constant TPI | 0.467<br>0.119 | 0.736<br>0.002 | <br>0.747 | 0.634<br>5.385 | 0.532<br>0.000 |

Note that the model summary indicates that the variance accounted for by TPI scores is 0.558, or about 56%. The ANOVA table indicates that this is a significant amount of variance (*F* of 28.997 with 1 and 23 degrees of freedom). Additionally, the beta value (standardized regression weight) in the coefficients table is equal to 0.747, the same as the zero-order correlation between TPI scores and number of work teams involved in. The *B* (unstandardized regression weight) associated with the TPI scores is 0.119 with a standard error of 0.002. This yields a *t* value of 5.385, and it is significant. The other *B* (0.467) is that of the intercept value in the regression equation.

Criterion studies can be conducted concurrently or predictively. There is no difference in the analysis or interpretation of the magnitude of the relationship between the variables using one or the other. The difference between the two methods is when the data for the criterion variable are collected. For concurrent studies, the predictor and criterion variables are collected at the same time. Thus, a study

*(Text continues on page 212)*

**Box 9.2** Hierarchical Regression of Number of Work Teams Regressed First on Age, and Then Age and TPI Scores Using SPSS

| *Model Summary* | | | | |
|---|---|---|---|---|
| *Model* | *R* | *R-Square* | *Adjusted R-Square* | *Std. Error of the Estimate* |
| 1 | 0.634 | 0.402 | 0.376 | 1.55824 |
| 2 | 0.749 | 0.560 | 0.520 | 1.36599 |

| *Change Statistics* | | | | |
|---|---|---|---|---|
| *R-Square Change* | *F Change* | *df1* | *df2* | *Sig. F Change* |
| 0.402 | 15.450 | 1 | 23 | 0.001 |
| 0.158 | 7.930 | 1 | 22 | 0.010 |

| *Coefficients* | | | | | |
|---|---|---|---|---|---|
| *Model* | *Unstandardized Coefficients* | | *Standardized Coefficients* | *t* | *Sig.* |
| | *B* | *Std. Error* | *Beta* | | |
| 1 (Constant) | 9.439 | 1.379 | | 6.85 | 0.000 |
| Age | –0.156 | 0.040 | –0.634 | –3.93 | 0.001 |
| 2 (Constant) | 1.542 | 3.054 | | 0.505 | 0.619 |
| Age | –0.0215 | 0.059 | –0.087 | –0.363 | –0.720 |
| TPI | 0.108 | 0.038 | 0.676 | 2.816 | 0.010 |

Note that in the Model Summary, for Model 1 overall, age accounts for 40.2% of the variance in number of work teams. For Model 2 overall, age and TPI scores account for 56.0% percent of the variance in number of work teams. The change statistics indicate that Model 1 significantly changes the variance accounted for from 0% to 40.2% (*F* of 15.450, sig. 0.001). Model 2 adds an additional 15.8% variance and it, too, is significant (*F* of 7.930, sig. 0.010). The coefficients assess both Models 1 and 2 separately. For Model 1, we can see that age is negatively related to number of work teams ($B = -0.156$) and it is significant ($t = -3.93$, sig. 0.001). Examining the coefficients for Model 2, we can see that TPI scores are significant predictors of number of work teams above and beyond age ($t = 2.816$, sig. 0.010), but age does not significantly predict number of work teams over and above TPI scores ($t = -0.363$, sig. 0.720).

**Table 9.4** Data Set for Simultaneous Multiple Regression of Performance on Cognitive Skill, Honesty, Conscientiousness, and TPI

| *Case* | *Cognitive Skill* | *Honesty* | *Conscientiousness* | *TPI* | *Performance* |
|---|---|---|---|---|---|
| 1 | 20 | 1 | 6 | 10 | 1 |
| 2 | 25 | 2 | 5 | 12 | 2 |
| 3 | 26 | 3 | 4 | 14 | 3 |
| 4 | 24 | 1 | 3 | 42 | 2 |
| 5 | 26 | 2 | 7 | 50 | 3 |
| 6 | 27 | 3 | 8 | 16 | 2 |
| 7 | 30 | 4 | 9 | 29 | 3 |
| 8 | 33 | 3 | 12 | 27 | 2 |
| 9 | 35 | 5 | 13 | 26 | 3 |
| 10 | 26 | 4 | 3 | 25 | 4 |
| 11 | 25 | 3 | 5 | 30 | 3 |
| 12 | 24 | 2 | 6 | 33 | 4 |
| 13 | 28 | 5 | 18 | 24 | 3 |
| 14 | 23 | 5 | 16 | 25 | 3 |
| 15 | 24 | 6 | 14 | 26 | 2 |
| 16 | 25 | 9 | 20 | 30 | 4 |
| 17 | 28 | 8 | 3 | 35 | 5 |
| 18 | 30 | 3 | 4 | 34 | 3 |
| 19 | 35 | 6 | 5 | 18 | 5 |
| 20 | 36 | 5 | 20 | 14 | 4 |
| 21 | 34 | 9 | 19 | 20 | 5 |
| 22 | 37 | 4 | 18 | 22 | 4 |
| 23 | 38 | 7 | 4 | 23 | 5 |
| 24 | 39 | 6 | 6 | 19 | 4 |
| 25 | 35 | 7 | 8 | 27 | 5 |
| 26 | 40 | 3 | 4 | 26 | 4 |

*(Continued)*

**Table 9.4** (Continued)

| *Case* | *Cognitive Skill* | *Honesty* | *Conscientiousness* | *TPI* | *Performance* |
|---|---|---|---|---|---|
| 27 | 41 | 6 | 6 | 35 | 5 |
| 28 | 40 | 2 | 3 | 40 | 6 |
| 29 | 42 | 7 | 8 | 41 | 4 |
| 30 | 43 | 4 | 9 | 42 | 5 |
| 31 | 44 | 8 | 9 | 43 | 6 |
| 32 | 45 | 9 | 10 | 46 | 5 |
| 33 | 43 | 4 | 12 | 47 | 6 |
| 34 | 45 | 9 | 13 | 48 | 7 |
| 35 | 46 | 4 | 13 | 43 | 5 |
| 36 | 47 | 5 | 14 | 45 | 7 |
| 37 | 47 | 7 | 15 | 46 | 6 |
| 38 | 46 | 3 | 20 | 47 | 6 |
| 39 | 45 | 8 | 16 | 48 | 5 |
| 40 | 42 | 9 | 17 | 39 | 6 |
| 41 | 41 | 6 | 14 | 38 | 7 |
| 42 | 35 | 5 | 13 | 37 | 8 |
| 43 | 39 | 8 | 12 | 47 | 7 |
| 44 | 38 | 4 | 18 | 45 | 8 |
| 45 | 37 | 6 | 18 | 46 | 8 |
| 46 | 40 | 3 | 17 | 36 | 7 |
| 47 | 47 | 8 | 16 | 38 | 9 |
| 48 | 48 | 8 | 20 | 41 | 6 |
| 49 | 49 | 3 | 20 | 42 | 7 |
| 50 | 49 | 6 | 19 | 49 | 8 |

**Box 9.3** Simultaneous Solution Multiple Regression Using Cognitive Skills, Honesty, Conscientious Personality, and TPI Scores to Predict Job Performance Using SPSS

| *Model Summary* | | | | |
|---|---|---|---|---|
| *Model* | *R* | *R-Square* | *Adjusted R-Square* | *Std. Error of the Estimate* |
| 1 | 0.801 | 0.641 | 0.609 | 1.20768 |

| *ANOVA* | | | | | |
|---|---|---|---|---|---|
| *Model* | *Sum of Squares* | *df* | *Mean Square* | *F* | *Sig.* |
| Regression | 117.088 | 4 | 29.272 | 20.070 | 0.000 |
| Residual | 65.632 | 45 | 1.458 | | |
| Total | 182.720 | 49 | | | |

| *Coefficients* | | | | | |
|---|---|---|---|---|---|
| | *Unstandardized Coefficients* | | *Standardized Coefficients* | | |
| *Model* | *B* | *Std. Error* | *Beta* | *t* | *Sig.* |
| (Constant) | –1.678 | 0.770 | | –2.18 | 0.035 |
| Cognitive | 0.106 | 0.029 | 0.460 | 3.68 | 0.001 |
| Honesty | 0.105 | 0.083 | 0.127 | 1.26 | 0.213 |
| Conscien. | 0.0453 | 0.034 | 0.136 | 1.35 | 0.185 |
| TPI | 0.0483 | 0.019 | 0.284 | 2.51 | 0.016 |

Using a simultaneous (direct) solution, a number of indicators can be assessed at once regarding each one's usefulness in accounting for overall and unique variance in job performance. Note that the model using all four variables as predictors of job performance allows us to account for 64% of the variance in job performance scores. The ANOVA table indicates that this is significantly different from 0% ($F = 20.070$, sig. 0.000). In the coefficients table, we can see that the only variables that predict job performance above and beyond that of any of the other three predictors are cognitive skills ($t = 3.68$, sig. 0.001) and TPI ($t = 2.41$, sig. 0.016). This is also demonstrated in the magnitude of the beta values (0.460 for cognitive skills and 0.284 for TPI). Honesty (0.136) and conscientious personality (0.136) are relatively less useful in predicting job performance.

where both TPI scores and number of teams worked on in the past two years is collected at the same time from a group of employees would be a concurrent criterion-related study. If TPI scores on a group of employees is collected at Time 1, and then, after two years go by, that same group of employees is asked how many teams they had worked on, then the criterion data are gathered at Time 2. This is called a predictive criterion-related study.

There are *post*dictive studies, where one gathers a criterion variable from the past. For example, a group of workers might be asked how many teams or clubs they were involved in during their high school years. Then they would be asked to complete the TPI scale, and their scores would be used to predict, retroactively, the number of teams or clubs those individuals were involved in during their high school years.

One issue that comes up in criterion-related studies is that, if there are multiple predictors, should some of them be weighted more than others when making a decision? For example, if a test of cognitive ability and test of a work sample are both used to predict job performance, should one be given more weight in the decision-making process than the other? The evidence suggests that using complex weighting schemes for the predictors does not offer much improvement over not doing so (i.e., unit weighting; Aamodt & Kimbrough, 1985).

*Convergent/Divergent Assessment.* Validation via convergent and divergent assessment was first introduced by Campbell and Fiske (1959) almost 50 years ago. This process has to do with relationships between the construct of interest and other similar or dissimilar constructs. For example, it would be expected that TPI scores should be somewhat positively related to other variables that assess sociability. It would be expected that TPI scores would be negatively related to variables such as independence or autonomy.

As a concrete example, suppose TPI scores were correlated with social interaction and social relation values scores (Macnab, Fitzsimmons, & Casserly, 1987; as was done in Kline, 1999). There should be a significant amount of shared variance between the constructs. This provides convergent information that the TPI construct is indeed assessing what it purports to assess. However, if the correlations were very high (0.80 or more), then the TPI construct would be considered to be too redundant with social interaction and social relation. To ensure that the construct diverges enough from the others to be considered unique, the correlations between such similar constructs should be moderate (between 0.30 and 0.50).

In addition, it is also expected that correlations of the similar constructs (social interaction and social relation) with other criteria (e.g., team cohesion) acted similarly in terms of strength and direction as did TPI scores. Again, this helps to provide convergent evidence about the validity of the TPI construct.

*Upper Bounds of Validity and Correction for Unreliability.* As just reviewed, criterion-related validity studies depend on assessing the relationship between a test (predictor) and outcome (criterion), usually with a correlation or regression coefficient. Both the predictor and criterion should be assumed to be fallible (unreliable) measures. Almost no test (predictor) has a reliability coefficient of 1.0, and the outcomes (criteria) are often even more plagued by unreliability. While test developers and users are quite

cognizant of the issues associated with the reliability of the predictor variable, the criterion may not even be subjected to any type of reliability assessment. This "criterion problem" was cited many years ago (Ghiselli, 1956), but continues to plague those using criterion measures for purposes of test score validation (e.g., Binning & Barrett, 1989).

While this may seem to be of only theoretical importance, note that the maximum value of a validity coefficient ($r_{xy}$) is

$$r_{xy} = \sqrt{(r_{xx})(r_{yy})}, \tag{9–1}$$

where $r_{xx}$ is the reliability of the *X* variable and $r_{yy}$ is the reliability of the *Y* variable. That is, the upper limit of any criterion-related validity coefficient, $r_{xy}$, is equal to the square root of the product of the reliabilities of the predictor ($r_{xx}$) and criterion ($r_{yy}$). For example, if test scores and criterion scores with reliabilities of $r_{xx} = 0.50$ and $r_{yy} = 0.50$, respectively, are correlated with each other, the upper limit of their relationship (the validity coefficient) is

$$\begin{aligned} r_{xy} &= \sqrt{(0.50)(0.50)}, \\ &= \sqrt{0.25}, \\ &= 0.50. \end{aligned}$$

If the predictor and criterion are much more reliable, let's say $r_{xx} = 0.90$ and $r_{yy} = 0.80$, the upper limit of their relationship is

$$\begin{aligned} r_{xy} &= \sqrt{(0.90)(0.80)}, \\ &= 0.85. \end{aligned}$$

The lower the reliability of either, the less likely it is that a validity coefficient will be significant. Therefore, the practical implications of the relationship between reliability and validity are extremely important. This is the case because, even at the best of times, the validity coefficient will not get close to its upper limit.

So, what some individuals do to get around this issue is engage in a process called *correction for reliability attenuation.* This means that first an observed validity coefficient is calculated. Then the question is asked: What would the corrected validity coefficient be if the predictor, criterion, or both were perfectly reliable? For example, if $r_{xy} = 0.28$, $r_{xx} = 0.70$, and $r_{yy} = 0.80$, what would the corrected $r_{x'y}$ be if $r_{xx} = 1.00$ (i.e., $r_{xt}$)? The formula for correcting the validity coefficient for unreliability in the predictor is

$$\begin{aligned} r_{x'y} &= r_{xy}/\sqrt{r_{xx}} \\ &= 0.28/\sqrt{0.70}, \\ &= 0.28/0.84, \\ &= 0.33. \end{aligned} \tag{9–2}$$

Thus, if it could be assumed that the test scores were perfectly reliable, then the validity coefficient would go up from 0.28 to 0.33. The formula for correcting the validity coefficient for unreliability in the criterion is

(9–3)
$$r_{xy'} = r_{xy}/\sqrt{r_{yy}}$$
$$= 0.28/\sqrt{0.80},$$
$$r_{xy'} = 0.31.$$

Thus, if could be assumed that the criterion values were perfectly reliable, then the validity coefficient would go up from 0.28 to 0.31. Now assume that both predictor and criterion were corrected for attenuation. This formula for this correction is

(9–4)
$$r_{x'y'} = r_{xy}/\sqrt{(r_{xx})(r_{yy})},$$
$$= 0.28/\sqrt{(0.70)(0.80)},$$
$$= 0.37.$$

Now the validity coefficient has jumped from 0.28 to 0.37! Be cautious in interpreting such corrected values. Note that the more unreliable the variable is to start with, the more this "correction" will boost the validity coefficient.

It is more reasonable to correct the predictor for unreliability than it is to correct the criterion. This is because test content can be changed to enhance its reliability. It is less defensible to correct the criterion for unreliability unless it, too, can be made more reliable.

*Range Restriction and Correction.* One common problem in correlational studies is that one or both of the variables have a *restricted range.* For example, assume I was interested in seeing the relationship between accident rates and driver's scores on a paper-and-pencil driving test to predict accidents during the first year of driving. If the driver's test can be assumed to be useful, there should be a negative relationship between test scores and accident rates. Assume that a data set based on 30 individuals is collected. All of them were given the test and all of them were given a driver's license regardless of their test score. The data for this example is shown in Table 9.5.

If a correlation analysis is run on the data, the resulting correlation is –0.72, which is very high. The scatterplot of the data is shown in Figure 9.1. However, if a decision rule had been invoked where only those with driving test scores of 80 or more pass and get a license, then those scoring less than 80 would not have been on the roads and getting into accidents. The correlation based only on those who did pass would be very low ($r = 0.13$) and nonsignificant. The scatterplot of these scores is shown in Figure 9.2.

This is a rather extreme example. However, it is not uncommon to have restricted ranges in the data. If either or both of the variables to be correlated have restricted ranges, be cautious in assuming the relationship between them is nonsignificant.

There is a correction one can make to estimate what the correlation would be if the range was not restricted. If it is to be used, the population variance needs to be known or estimated in advance. Often this can be determined by using information contained in past studies or by using the distributions calculated for the $r_{wg}$ in Chapter 8. In the driving test example, assume that the 30 cases actually represented

**Table 9.5** Data for Assessing the Relationship Between Driver's Test Scores and First-Year Accident Rates

| *Case* | *Test Score* | *Accidents in First Year Driving* |
|---|---|---|
| 1 | 50 | 2 |
| 2 | 60 | 1 |
| 3 | 45 | 2 |
| 4 | 58 | 2 |
| 5 | 62 | 1 |
| 6 | 72 | 1 |
| 7 | 70 | 2 |
| 8 | 66 | 1 |
| 9 | 83 | 1 |
| 10 | 77 | 2 |
| 11 | 76 | 1 |
| 12 | 72 | 1 |
| 13 | 55 | 3 |
| 14 | 47 | 3 |
| 15 | 83 | 1 |
| 16 | 92 | 1 |
| 17 | 67 | 2 |
| 18 | 77 | 1 |
| 19 | 95 | 1 |
| 20 | 87 | 1 |
| 21 | 88 | 0 |
| 22 | 85 | 0 |
| 23 | 79 | 1 |
| 24 | 86 | 0 |
| 25 | 91 | 0 |
| 26 | 90 | 1 |
| 27 | 87 | 0 |
| 28 | 76 | 2 |
| 29 | 70 | 1 |
| 30 | 72 | 2 |

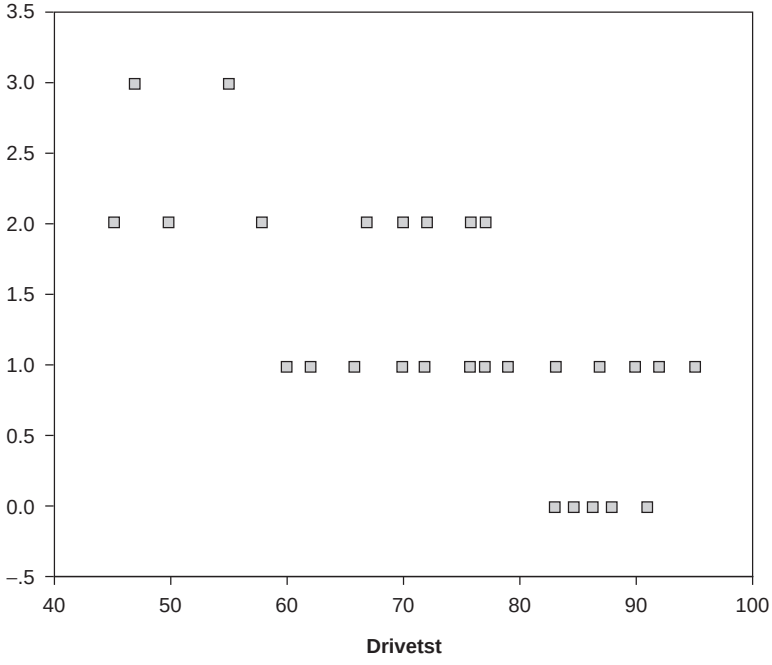


**Figure 9.1** Scatterplot of Driving Test Scores and Accidents

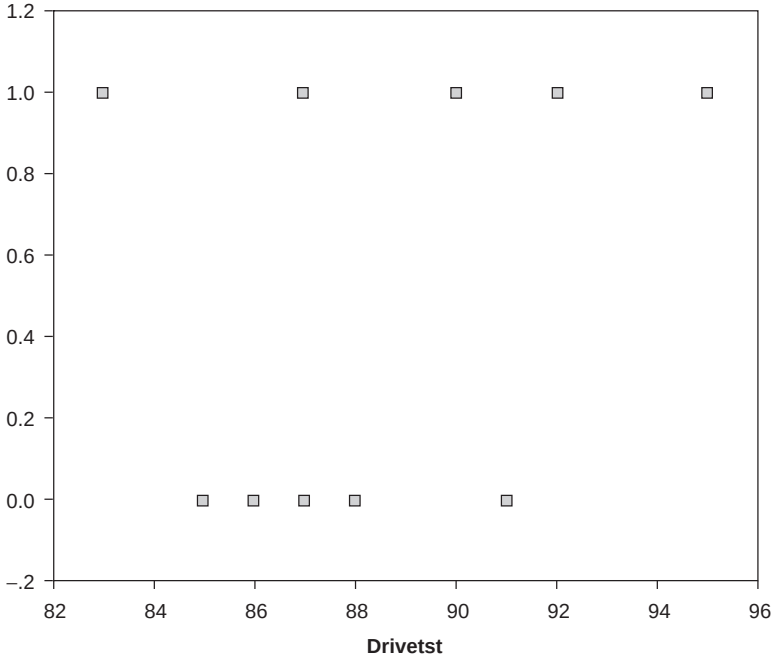


**Figure 9.2** Scatterplot of Driving Test Scores and Accidents With a Restricted Range

the population. In this case, the standard deviation of their scores on the driving test scores was 13.87. When the sample was restricted to only those scoring 80 or more on the test, the standard deviation of these individuals was 3.78. The correction for attenuation due to range restriction formula is given by Guilford and Fruchter (1973):

$$r_{xy'} = [r_{xy}(\sigma_{pop}/\sigma_{sample})]/\sqrt{1 - r_{xy}^2 + r_{xy}^2 (\sigma_{pop}^2/\sigma_{sample}^2)}, \quad (9–5)$$

where $r_{xy'}$ = the corrected validity coefficient, $r_{xy}$ = the calculated validity coefficient, $\sigma_{pop}$ = the population standard deviation, and $\sigma_{sample}$ = the sample standard deviation.

So, substituting in the driving test example,

$$\begin{aligned} r_{xy'} &= [0.13(13.87/3.78)]/\sqrt{(1 - 0.0169) + 0.0169\ (192.38/14.29)}, \\ &= 0.48/\sqrt{0.9831 + 0.2275}, \\ &= 0.48/1.10, \\ &= 0.43. \end{aligned}$$

The estimate of the correlation between driving test and accident rates based on the cases who scored 80 or more on the test and corrected for range restriction is 0.43. This is substantially higher than 0.13, so the restricted range on the driving test variable clearly attenuated the correlation. Thus, it is very important to assess both predictor and criterion variables for range restriction and take steps to deal with the problem if it is appropriate to do so.

*Sample Size.* Sample sizes are sometimes a problem in criterion-related validity studies. A simple examination of a table to test the significance of a correlation coefficient will drive home this point. For example, if the relationship between GRE test scores and graduate school performance is calculated, there may only be 10 students to supply the data for the analysis if one department was used. A validity coefficient of 0.632 will be needed to be considered statistically significant with that number of cases. However, assume that information on graduate students was pooled across several departments; by doing so, the analysis could be conducted on 30 students. Now a validity coefficient of only 0.361 is needed to be considered statistically significant. Assume a sample size of more than 100 cases. It will be possible to have a statistically significant relationship if the validity coefficient is less than 0.20.

Thus, it is important in interpreting studies that report validity coefficients to examine both the size and representativeness of the sample on which the data were collected.

*Making Decisions: Cutoffs and the Roles of the Standard Error of Estimate and the Standard Error of Measurement.* The outcomes of criterion-related validity studies are often used to set standards by invoking a cutoff score below which individuals are not selected. For example, assume that a criterion-related validity coefficient suggests that there is a significant and positive relationship between children's IQ

scores and success in a gifted scholastic program. The relationship is strong enough that IQ scores are now going to be used to make decisions about which children to allow into the program and which to exclude. How strong is "strong enough" has to do with the standard error of estimate. The standard error of estimate was briefly introduced in Chapter 1 as part of the output of a regression analysis, and it plays a large role in the accuracy of predicting criterion scores. An example will assist in demonstrating the centrality of the standard error of estimate.

Assume that in a pilot study of the usefulness of IQ scores in predicting performance in gifted programs, 100 students with a wide range of IQ scores were all allowed into a gifted program. The criterion variable was the teacher's rating of the success of each student on a 10-point scale where 4 was the cutoff for "successful completion."

Now, IQ scores are regressed on ratings of success. Assume the calculated $R^2$ value was 0.35, indicating that 35% of the variance in success is predicted by IQ scores. Further assume that the calculated standard error of estimate was 0.50. The standard error of estimate is equal to the standard deviation of the criterion score (success in this example) for individuals with the same score on the predictor (IQ in this example).

So, assume that students with an IQ score of 120 have a predicted success score of 2.5. The standard error of estimate can be used to set a confidence interval around the predicted score of 2.5. To set the 95% confidence interval, the relevant $t$ value is needed. Because a 95% confidence interval is desired, the $\alpha$ is 0.05, and because there were 100 participants, the degrees of freedom are 98 (i.e., $100 - 2$). The $t$ value at $\alpha = 0.05$ with 98 degrees of freedom is equal to 1.99. Then the standard error of estimate is multiplied by that $t$ value ($0.50 \times 1.99 = 0.995$). Finally, the confidence interval is set by subtracting 0.995 from and adding 0.995 to the estimated score (i.e., $2.5 - 0.995 = 1.505$ and $2.5 + 0.995 = 3.495$).

It is concluded that an individual with an IQ of 120 will have a predicted success score between 1.505 and 3.495 with 95% confidence. In more practical terms for this example, those with IQ scores of 120 or less would not meet the success criterion of 4 because the highest they would be expected to obtain on the criterion would be 3.495, 95 times out of 100. The IQ score where the 95% confidence interval upper bound includes the success criterion (4 in this example) could appropriately be used as the cutoff score.

This is a simplification of the process by which cutoff scores are set. It should be obvious that in such pilot studies, the larger and more representative the sample, the better the future decisions will be. Unrestricted range in the predictor and excellent training of the teachers in making their criterion ratings would also be critical to the process being above reproach. In such high stakes decision making, the process for setting cutoffs is always under scrutiny. Often, several different pilot studies of the type described are carried out and the results pooled and interpreted. Ultimately, a judgment call is made about where to set the cutoff, but it should be based on the results of methodologically and statistically sound studies.

In Chapter 7, the standard error of measurement was used to set confidence intervals around an estimated true score for any one individual. This process is

particularly salient when the scores are being used for decision making. Continuing with the IQ and gifted program example, assume for the sake of argument that, based on a series of studies, it has been shown that children must score statistically above 125 on a measure of IQ to be successful, and therefore that is the test cutoff score to be used before being admitted into the program. Now suppose there is a child seeking admission, and she has an obtained test IQ score of 125.

Let's assume that the population mean and standard deviation for this test are 100 and 15, respectively. Further assume that the reliability of the test is 0.95, and therefore her estimated true score is 118.75 (i.e., $125 \times 0.95$). Further assume that the standard error of measurement for the IQ test is 3.35. The 95% confidence interval around her score, then, is

$$\begin{aligned} &118.75 + (1.96 \times 3.35), \\ &\quad = 118.75 + 6.57, \\ &\quad = 112.18 <\!\!-\!\!-\!\!> 125.32. \end{aligned}$$

Now the question is a practical one: How much "above 125" does the score have to be? In this case, it is above 125 (by 0.32) but it rounds to 125. Interpretation of such differences can become a legal nightmare. This is why it is critical that more than just one indicator be used to make such decisions. It would not be appropriate to base the judgment on a single test score alone. Other pieces of information that might be useful in this example would be test scores on classroom work, portfolios of work samples, interviews with teachers, and so forth. Then the preponderance of the evidence can be used to make a final decision about the appropriateness of admission.

There are situations where cutoff scores are the only indicator used (e.g., licensing examinations for professionals), and therefore information about the standard error of estimate, test reliability, and standard error of measurement becomes essential in defending decisions based on test scores.

## Multiple Criteria

To this point, it has been assumed that the criterion is a single variable. However, most variables of interest are multivariate. Job performance, scholastic performance, social adjustment, marital satisfaction, and so forth have several facets, or elements. A question arises from this phenomenon: How should facet information be combined to come up with a single criterion?

The process of dealing with this issue often juxtaposes a multiple hurdle/multiple cutoff model versus a compensatory model. It will depend on the context as to which makes most sense. For example, if the criterion is the skill set needed to be a pizza delivery person, then there might be several predictors: (a) a valid driver's license so that the person can deliver the pizza, (b) quantitative skills for collecting the correct amount owed, (c) map-reading skills for getting around the city, and (d) interpersonal skills for interacting with the customer in the hope

of obtaining repeat business. Tests might be used for some of these predictors. In the end, it really is up to the pizza store managers to decide how to weight these various facets.

Some of the facets may be critical and stop or continue the process (e.g., if the job applicant does not have a valid driver's license, then the individual cannot do the job). Other facets may be more difficult to decide upon, and some sort of decision process needs to be used where facets b, c, and d described above are differentially weighted as 2, 2, and 1, respectively. This gives twice as much weight to the quantitative and map-reading skills than to interpersonal skills. However, these more complex weighting schemes are not always desirable, and it is recommended that when several criteria are combined, simple unit weighting is actually a better approach (Society for Industrial and Organizational Psychology, 2003).

Whether criteria facets are differentially weighted or not, the decision of doing so or not must be defensible. Again, SMEs are invaluable in making these determinations. In employment or other high stakes situations, the process by which the criterion is developed is as important as the test development. Unfortunately, not nearly as much care goes into the development of sound criteria as goes into sound test development.

In the end, like it or not, a decision must be made, and therefore the information is combined somehow. Such decisions include hire or not, promote or not, give a license or not, stream a student or not, and deliver a course of treatment or not. Multiple pieces of information will go into such overall decisions and those making the decisions will often invoke their own decision rules for doing so. The more solid statistical information that can be provided to these decision makers about how to combine the information and the better the predictors, the more likely their decisions are going to be of high quality.

## Classification Approaches to Test Score Validation

Sometimes it is not of interest to be as fine-grained as many of the criterion-related validity studies described thus far allow for. For example, from a practical perspective, the specific score an individual gets on a criterion measure may not be as important as whether the individual succeeds or meets expectations. In such cases, classification approaches to assessing criterion-related validity studies are appropriate.

*Predictive Accuracy With a Dichotomous Predictor and Criterion.* When one dichotomous predictor (e.g., pass/fail) and one dichotomous criterion (e.g., succeed/fail) form the data set to be analyzed, they can be placed in a $2 \times 2$ matrix such as those used in signal detection theory (see Table 9.6). Here, A = hits (pass on the predictor and successful on the criterion), B = correct rejections (fail on the predictor and unsuccessful on the criterion), C = false alarms (pass on the predictor and unsuccessful on the criterion), and D = misses (fail on the predictor

**Table 9.6** Predictive Accuracy When the Predictor and Criterion Are Dichotomized

| | | |
|---|---|---|
| Successful | D<br>Misses | A<br>Hits |
| Unsuccessful | B<br>Correct Rejections | C<br>False Alarms |
| | Fail | Pass |

and successful on the criterion). False alarms can be very costly because if one hires or accepts someone but that person turns out to be unsuccessful on the job, then recruiting costs, training time, salary, and so forth are lost. Misses are also expensive in terms of lost opportunities. Good predictors will maximize the number of individuals in cells A and B while minimizing the number in cells C and D.

It is clear from this table that the stronger the relationship between predictor and criterion, the better the predictive accuracy. In this case, more of the cases will cluster around an imaginary line that cuts across the square diagonally. That means fewer cases will be in the C and D quadrants. In addition, given that all else remains constant, one can shift around the cutoff score on the predictor to change the numbers in the quadrants. So, if the line was moved from the center to the left, more cases would be in the A and C quadrants. If the line was moved to the right, more cases would be in the B and D quadrants. Finally, when the successful/unsuccessful split is 50%/50%, the predictor is most powerful (i.e., has the opportunity to be of most value). To understand this concept, visualize an example where there are very few cases in the A and D quadrants or, alternatively, where there are very few cases in the B and C quadrants. If this was the case, then correct decisions (hits and correct rejections) would be harder to make.

The base rate for success can only be determined if one were to not use any predictors at all. For example, suppose 100% of the 50 applicants for being a registered psychologist over the next six months were licensed. Then, a year later, it is determined that 50% of them were unsuccessful (by what means *unsuccessful* was determined is not relevant to this discussion, although it would be, of course, in another context). Then the question can be asked: What would be the improvement in the accuracy of judgments (increases in quadrants A and C) if a predictor such as "applicants have to pass an examination" is added to the decision process? To answer this, another group of 100 applicants is given the test, but only those who pass the test (i.e., obtain a score of 70% or more) are licensed ($n = 80$), and then it is determined a year later how many were unsuccessful. Assume, in this example, that 25% were unsuccessful. Is the 25% reduction in unsuccessful psychologists significant?

The test for differences between two proportions can be used to answer this question:

$$z = (P_{\text{success1}} - P_{\text{success2}}) / \sqrt{(\overline{p})(1 - \overline{p})(\tfrac{1}{n_1} + \tfrac{1}{n_2})}, \tag{9–6}$$

where $P_{\text{success1}}$ = the proportion of successes in sample 1, $P_{\text{success2}}$ = the proportion of successes in sample 2, $n_1$ = the number of cases in sample 1, $n_2$ = the number of cases in sample 2, and $\overline{p}$ = the pooled estimate of the population proportion of successes, $\overline{p} = (X_1 + X_2)/(n_1 + n_2)$, where $X_1$ and $X_2$ are the number of successes in samples 1 and 2, respectively.

In the example, $n_1 = 50$, $n_2 = 80$, $P_{\text{success1}} = 0.50$, $P_{\text{success2}} = 0.75$. Thus, $\overline{p} = (25 + 60)/(50 + 80)$, $\overline{p} = 85/130 = 0.65$. Now the question can be answered:

$$\begin{aligned} z &= (0.50 - 0.75)/\sqrt{(0.65)(0.35)(0.02 + 0.0125)}, \\ &= (-0.25)/\sqrt{(0.65)(0.35)(0.0325)}, \\ &= -0.25/0.086, \\ &= -2.91. \end{aligned}$$

The significance level that needs to be exceeded for the proportional difference to be considered significant is ± 1.96. Because −2.91 does exceed this value, it can be concluded that there is a significant improvement in the proportion of successful psychologists after implementing the new test.

In a selection context (whether it be for employment, graduate school, educational programming, treatment programs, etc.) where the relationship between the predictor and criterion is known, the success base rate is known, and the selection ratio (proportion taken in versus rejected) is known, Taylor-Russell (1939) tables can be used to determine the improved accuracy by changing the selection ratio. For example, assume that the success base rate for students to survive the first year of medical school is 0.60, and we want to improve that by instituting a test as a selection device that has been used in other medical schools. Studies have shown that the validity coefficient between the test and medical school success in the first year is about 0.40. It is decided that in the upcoming year, 20% of the applicants will be selected (selection ratio of 0.20). Using the Taylor-Russell tables (example is in Table 9.7), it is determined that the proportion of students who will be successful the first year will now be 0.81. This is an improvement of 21% (0.81 − 0.60) and is likely to be of practical importance given the cost of educating students in medical schools.

*Discriminant Function Analysis.* It is possible to have multiple predictors, some of them continuous and others dichotomous, for a given criterion of success. If this is the case, then an assessment of predictive accuracy can be carried out using discriminant function analysis (DFA). DFA answers two questions: (a) Do the predictors help to classify individuals as successful or not better than chance? and (b) Which of the predictors are most relevant in that prediction?

**Table 9.7** Taylor-Russell Table: Base Rate = 0.60

| | *Selection Ratio* | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *r* | .05 | .10 | .20 | .30 | .40 | .50 | .60 | .70 | .80 | .90 | .95 |
| .00 | .60 | .60 | .60 | .60 | .60 | .60 | .60 | .60 | .60 | .60 | .60 |
| .05 | .64 | .63 | .63 | .62 | .62 | .62 | .61 | .61 | .61 | .60 | .60 |
| .10 | .68 | .67 | .65 | .64 | .64 | .63 | .63 | .62 | .61 | .61 | .60 |
| .15 | .71 | .70 | .68 | .67 | .66 | .65 | .64 | .63 | .62 | .61 | .61 |
| .20 | .75 | .73 | .71 | .69 | .67 | .66 | .65 | .64 | .63 | .62 | .61 |
| .25 | .78 | .76 | .73 | .71 | .69 | .68 | .66 | .65 | .63 | .62 | .61 |
| .30 | .82 | .79 | .76 | .73 | .71 | .69 | .68 | .66 | .64 | .62 | .61 |
| .35 | .85 | .82 | .78 | .75 | .73 | .71 | .69 | .67 | .65 | .63 | .62 |
| .40 | .88 | .85 | **.81** | .78 | .75 | .73 | .70 | .68 | .66 | .63 | .62 |
| .45 | .90 | .87 | .83 | .80 | .77 | .74 | .72 | .69 | .66 | .64 | .62 |
| .50 | .93 | .90 | .86 | .82 | .79 | .76 | .73 | .70 | .67 | .64 | .62 |
| .55 | .95 | .92 | .88 | .84 | .81 | .78 | .75 | .71 | .68 | .64 | .62 |
| .60 | .96 | .94 | .90 | .87 | .83 | .80 | .76 | .73 | .69 | .65 | .63 |
| .65 | .98 | .96 | .92 | .89 | .85 | .82 | .78 | .74 | .70 | .65 | .63 |
| .70 | .99 | .97 | .94 | .91 | .87 | .84 | .80 | .75 | .71 | .66 | .63 |
| .75 | .99 | .99 | .96 | .93 | .90 | .86 | .81 | .77 | .71 | .66 | .63 |
| .80 | 1.0 | .99 | .98 | .95 | .92 | .88 | .83 | .78 | .72 | .66 | .63 |
| .85 | 1.0 | 1.0 | .99 | .97 | .95 | .91 | .86 | .80 | .73 | .66 | .63 |
| .90 | 1.0 | 1.0 | 1.0 | .99 | .97 | .94 | .88 | .82 | .74 | .67 | .63 |
| .95 | 1.0 | 1.0 | 1.0 | 1.0 | .99 | .97 | .92 | .84 | .75 | .67 | .63 |
| 1.00 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | .86 | .75 | .67 | .63 |

*Note*: From Taylor & Russell (1939).

DFA classifies individuals into their respective groups based on a linear combination of predictor variables. Like the regression equation, a DFA equation can be written as follows:

(9–7) $$D' = a + w_1 V_1 + w_2 V_2 + w_3 V_3 \cdots + w_x V_x,$$

where $D'$ = the discriminant function score for any individual, $a$ = the constant, $w_1$–$w_x$ = the weights assigned to each variable (predictor), and $V_1$–$V_x$ = the predictors.

The weights assigned to each predictor maximize the differences between the criterion groups. In a DFA, a series of discriminant functions will be generated from the computer program; however, not all of them will be significant. The maximum number that will be generated is equal to either (a) the number of criterion groups minus 1 or (b) the number of predictor variables, whichever is smaller. Each function must be orthogonal (uncorrelated) to the previous functions, and each subsequent function accounts for less and less variance in criterion. In the example used here for demonstration purposes, there will be two groups, successful and unsuccessful, so the maximum number of functions will be 2 – 1, or 1. It is worth noting here, though, that DFA can be used with a criterion that is made up of more than two groups.

In the first part of the DFA, there is a significance test for how well each function is able to place cases into the correct group (successful or unsuccessful). The test follows a $\chi^2$ distribution, and thus this statistic is used for assessing the significance of each function.

If the function is significant, then determining the contributions of each of the predictors (if there are more than one) is the next step. This information is found in the structure coefficients. The structure coefficients are equal to the zero-order correlations between the each of the predictors ($V_1$–$V_x$) and the discriminant function scores ($D'$).

Finally, the classification table provides an overall summary of the correct classification of individuals. The percentages on the diagonal are correct decisions and the off-diagonal percentages are incorrect decisions (i.e., the computer misclassified them).

In addition, a *jackknife,* or *cross-validation,* table is provided. To generate this table, each case one at a time is excluded from the analysis; a new DFA is generated and the excluded case is classified using the new discriminant function. Then a different case is excluded, another DFA run, and the case is classified using the new discriminant function. This proceeds until all the cases have been excluded and subsequently classified. In a way, it is like cross validating without having to gather data on another sample.

As an example, assume that we have a data set of 20 individuals; 10 of them are successful on the job and 10 are not. They are administered three tests to ascertain if their success is related to these instruments. The first is the TPI, the next is a cognitive skills test (IQ), and the third is an assessment of personal adaptability (ADAPT). A DFA analysis is run and the resulting discriminant function helps to classify the cases into "successful" and "unsuccessful" to a greater degree than chance (80% versus 50%). In addition, all of the predictors are useful in classifying the cases, although the TPI scores are most useful. Box 9.4 shows the output for this example using the discriminant program from SPSS.

## Group Differences and Test Bias

A potential problem using tests in high stakes environments is that the relationship between predictor and criterion act differently for identifiable subgroups.

*(Text continues on page 228)*

**Box 9.4** Discriminant Function Analysis: Using SPSS

In this analysis, a *direct* solution was requested, meaning that all three variables were forced into the analysis. Note, though, that it is possible to use a stepwise or hierarchical process in DFA.

**Table 9.8** DFA Group Statistics

| *Success* | *Mean* | *Standard Deviation* |
|---|---|---|
| 0 Unsuccessful | TPI = 20.3<br>IQ = 101.2<br>ADAPT = 4.2 | TPI = 4.19<br>IQ = 3.52<br>ADAPT = 1.81 |
| 1 Successful | TPI = 27.2<br>IQ = 105.9<br>ADAPT = 6.0 | TPI = 5.63<br>IQ = 8.90<br>ADAPT = 2.31 |

Table 9.8 shows the means for each group on the predictor variables. From the information, we see that the successful group is higher on all three scales than the unsuccessful. Our next question is whether these can be used to discriminate between the two groups.

**Table 9.9** DFA Eigenvalues

| *Function* | *Eigenvalue* | *% Variance* | *Cumulative %* | *Canonical Correlation* |
|---|---|---|---|---|
| 1 | 0.965 | 100.0 | 100.0 | 0.701 |

Note in Table 9.9 that there is only one function generated. The eigenvalue is a measure of shared variance (although the actual value of 0.965 is not relevant to us at this point). The percent variance is found by taking the eigenvalue for the function and dividing it by the total of all the eigenvalues for all the functions together. Because there is only one function, the percent variance is $0.965/0.965 \times 100$, or 100%. If there was more than one function generated, the relative contributions of each could be determined by examining this information. Because there is only one function, it is irrelevant. The canonical correlation represents the relationship between the discriminant function scores (*D'*) and the original grouping (successful vs. unsuccessful). The higher this is, the more the function separates the groups.

**Table 9.10** DFA Wilks's Lambda

| *Test of Function (s)* | *Wilks's Lambda* | *Chi-Square* | *df* | *Sig.* |
|---|---|---|---|---|
| 1 | 0.509 | 11.149 | 3 | 0.011 |

*(Continued)*

**Box 9.4** (Continued)

In Table 9.10, the function is tested for significance. The Wilks's lambda is an assessment of "badness of fit"—the lower the value the better (it ranges from 0–1). There is no distribution to test the Wilks's lambda for significance, and to do so we convert it to a chi-square (formula shown shortly). The chi-square of 11.149 with 3 degrees of freedom is significant. The degrees of freedom are equal to (# of groups on the criterion – 1) (# predictors); in our case, $1 \times 3 = 3$. If we had three groups and four predictors, we would have generated two functions before running out of degrees of freedom: the first would be (# groups – 1)(# predictors), or $2 \times 4 = 8$; the second would be (# groups – 2)(# predictors – 1), or $1 \times 3 = 3$. Degrees of freedom, then, are based on the number of groups and number of predictors. When either one becomes 0, there are no more degrees of freedom available to test the significance of the function. The significance level is 0.011, indicating that the function is significant in being able to correctly classify our 20 cases.

For completeness, the equation for converting the Wilks's lambda to chi-square is provided:

$$\text{Chi-square} = -[N - 1 - 0.5\,(p + q + 1)]\,[\log_n \text{Lambda}], \quad (9 - 8)$$

where $N$ = the sample size (20 in our case), $p$ = the number of criterion variables (always 1 in DFA), $q$ = the number of predictors (3 in our case), and $\log_n$ Lambda = the natural log of the Wilks's lambda value ($\log_n$ 0.509). So,

$$\begin{aligned}\text{chi-square} &= -[20 - 1 - 0.5\,(1 + 3 + 1)]\,[-0.6753],\\ &= -(19 - 2.5)(-0.6753),\\ &= -16.5 \times -0.6753,\\ &= 11.14.\end{aligned}$$

**Table 9.11** DFA Structure Coefficients Matrix

| | *Function 1* |
|---|---|
| TPI | 0.745 |
| ADAPT | 0.465 |
| IQ | 0.372 |

The cell entries in Table 9.11 are the zero-order correlations of each variable with the discriminant function scores (*D*'), and they help to determine which of the variables are most important in the function. These are always placed in descending order. There are no tests of significance for these coefficients, so we use a rule of thumb of 0.30 to determine if the coefficients are substantive. In this case, all three predictors contribute to the function, but the TPI scores do so more than the others.

**Table 9.12** DFA Canonical Discriminant Function Coefficients

| | *Function 1* |
|---|---|
| TPI | 0.178 |
| ADAPT | 0.060 |
| IQ | 0.196 |
| (Constant) | –11.415 |

Table 9.12 reports the unstandardized weights used to calculate the *D*′ scores. For example, if Case 1 had a score of 20 on TPI, 9 on ADAPT, and 100 on IQ, the *D*′ score, then, would be:

$$\begin{aligned} D' &= -11.415 + (0.178 \times 20) + (0.060 \times 9) + (0.196 \times 100), \\ &= -11.415 + 3.56 + 0.54 + 19.6, \\ &= -12.285 \end{aligned}$$

A *D*′ score for each case is created using this function, and these are used, then, for calculating the canonical correlation and the structure coefficients.

**Table 9.13** DFA Group Centroids

| *Success* | *Function1* |
|---|---|
| 0 | –0.932 |
| 1 | 0.932 |

The values in this matrix (Table 9.13) are equal to the mean average *D*′ values for each group. Thus, we can see that the average *D*′ score for those in the unsuccessful group was –0.932, and those in the successful group had mean *D*′ scores of 0.932. For each function, these sum to zero, although they are weighted by the sample size for each group. Because we had equal numbers in each group (10 in each), simply adding the centroids will give a value of 0. If there were two times as many cases in one group, the centroid for that group would be weighted twice as much before summing.

**Table 9.14** DFA Prior Probabilities for Groups

| *Success* | *Prior* |
|---|---|
| 0 | 0.50 |
| 1 | 0.50 |

*(Continued)*

**Box 9.4** (Continued)

Before calculating the classification statistics for each case, the prior probabilities of belonging to each group are reported (Table 9.14). Because we had equal numbers of cases in each group, the prior probability would be 0.50 for each. If there were 15 in one group and 5 in the other, the prior probabilities would be 0.75 and 0.25, respectively. You can also change the prior probabilities yourself in the command syntax. For example, if you knew in the population that prior probabilities were 0.33 and 0.67, then you could specify these values.

**Table 9.15** DFA Classification Results

| | | | *Predicted Group Membership* | | |
|---|---|---|---|---|---|
| | | *Success* | 0 | 1 | *Total* |
| Original | Count | 0 | 8 | 2 | 10 |
| | | 1 | 2 | 8 | 10 |
| | % | 0 | 80.0 | 20.0 | 100.0 |
| | | 1 | 20.0 | 80.0 | 100.0 |
| Cross Validated | Count | 0 | 7 | 3 | 10 |
| | | 1 | 2 | 8 | 10 |
| | % | 0 | 70.0 | 30.0 | 100.0 |
| | | 1 | 20.0 | 80.0 | 100.0 |

The first half of Table 9.15 (Original) shows the original group to which the cases belonged and the predicted group to which they belonged. We can see that 80% of the cases were correctly classified. This is 30% better than the prior chance probability of 50%. The second half of the table (Cross Validated) shows the results of the jackknife procedure. In this case, 75% of the cases were correctly classified. This half is helpful in letting you know to what extent you may be capitalizing on chance with the sample-specific characteristics. In this case, the cross validation shows that 75% of the cases would be correctly classified with another sample of similar size, which is still much better than chance.

Depending on the instrument, differences based on gender, race, language, culture, ethnicity, and so forth may or may not be expected. If test score differences emerge based on these types of demographic groups, it will limit the unrestricted use of norms generated for the test. When those subgroups are protected in the legal system, there is the opportunity for litigation issues to arise. The statistical issues and the relationship between predictor and criterion can provide some insight into whether or not the test is discriminatory.

**Table 9.16** Scores of 15 Males (Group = 1) and 15 Females (Group = 2) on Two Variables, *X* and *Y*

| *Group* | *X* | *Y* |
|---|---|---|
| 1 | 1 | 3 |
| 1 | 2 | 4 |
| 1 | 3 | 5 |
| 1 | 4 | 4 |
| 1 | 5 | 6 |
| 1 | 1 | 2 |
| 1 | 2 | 3 |
| 1 | 3 | 3 |
| 1 | 4 | 7 |
| 1 | 5 | 8 |
| 1 | 1 | 4 |
| 1 | 2 | 5 |
| 1 | 3 | 6 |
| 1 | 4 | 8 |
| 1 | 5 | 9 |
| 2 | 1 | 2 |
| 2 | 2 | 9 |
| 2 | 3 | 1 |
| 2 | 4 | 8 |
| 2 | 5 | 7 |
| 2 | 1 | 4 |
| 2 | 2 | 5 |
| 2 | 3 | 3 |
| 2 | 4 | 7 |
| 2 | 5 | 2 |
| 2 | 1 | 8 |
| 2 | 2 | 6 |
| 2 | 3 | 7 |
| 2 | 4 | 4 |
| 2 | 5 | 5 |

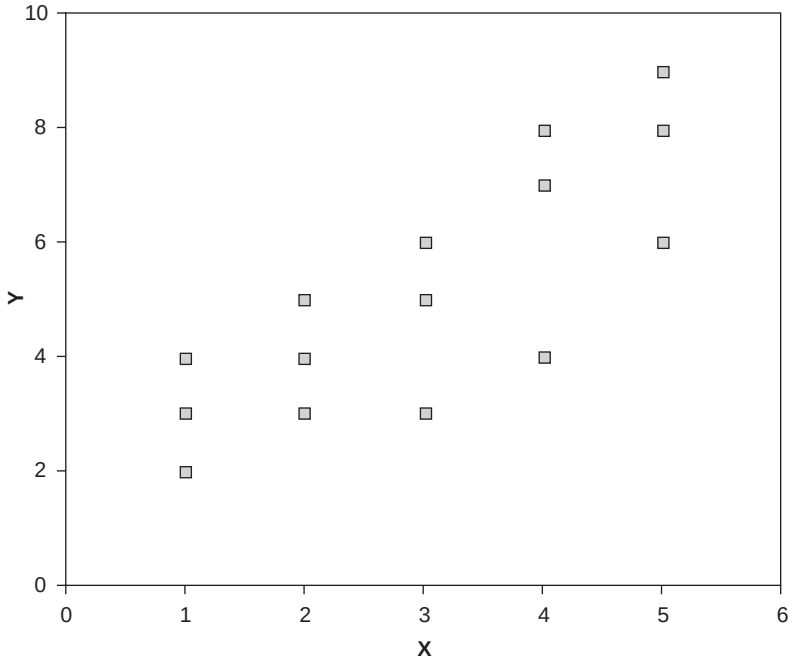


**Figure 9.3** Scatterplot of *X* and *Y* Scores for 15 Males

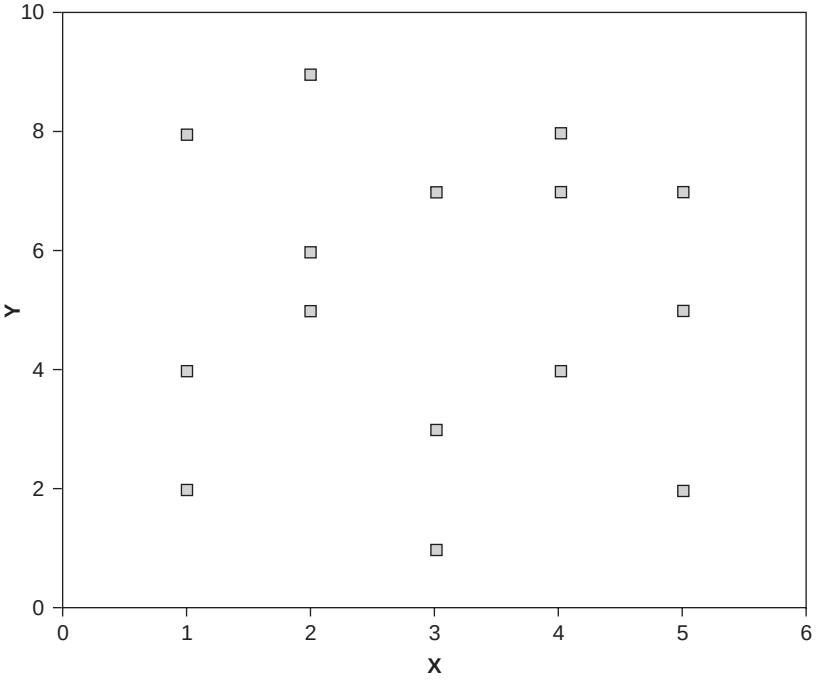


**Figure 9.4** Scatterplot of *X* and *Y* Scores for 15 Females

Assume the following example where the scatterplots (Figures 9.3 and 9.4) of two different groups look quite different, and notice what happens to the validity coefficient. The data for this example are given in Table 9.16; there are 15 males and 15 females. For males, the validity coefficient between *X* and *Y* is equal to 0.80. The validity coefficient for females is equal to –0.02.

Examining the scatterplots and the magnitude of the validity coefficients would lead us to conclude that *X* predicts *Y* for males but not for females. Take care to ensure that the test being used reports validity coefficient information on demographic subgroup differences. This will ensure that the scores are interpreted appropriately. In cases such as these, where there is a difference between the slopes of the regression lines for one group versus the other (which can be tested for statistical significance), differential validity is said to occur (Cleary, 1968) and this would be considered "unfair." This makes sense in the example; if the test (*X*) was used for females, there is no evidence to suggest that it predicts the criterion (*Y*), but if the test was used for males, there is evidence to suggest it predicts the criterion.

Groups can also differ only on the intercept. That is, the slopes of the regression equations are not statistically different from one another, but the lines for the two groups cross the y-axis in different locations (see Figure 9.5). In this case, the difference in criterion scores would be directly proportional to the difference in predictor scores, and the test would be considered unfairly used to predict the criterion. Specifically, the score on the criterion variable systematically differs at a given level of the predictor.

The line that is above the other in this example is the group that is "unfairly treated" because their predicted *Y* scores are consistently underestimated when

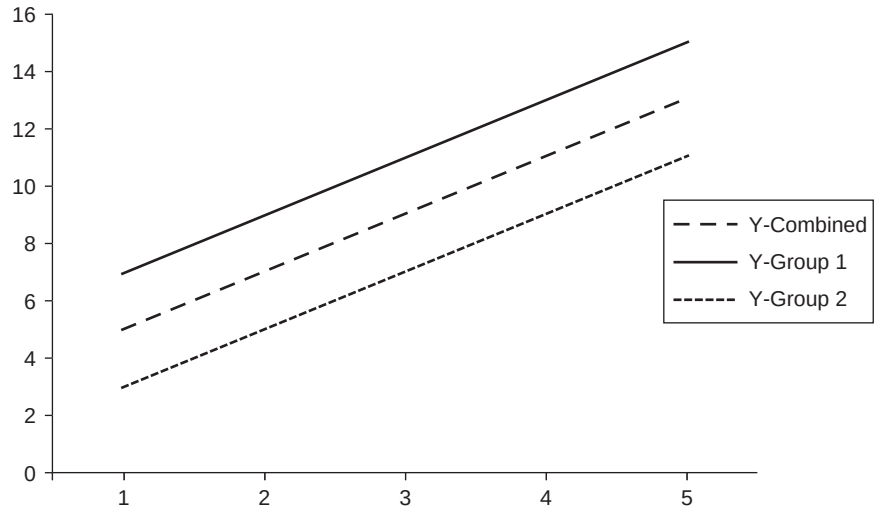


**Figure 9.5** Regression Lines With Two Different Intercepts for Focal and Reference Groups

using a regression line based on the entire sample, whereas those with the line below have consistently overestimated predicted scores. An example may clarify this idea. Assume that one regression equation is used for both groups:

$$Y' = 3 + 2.0(X).$$

If an individual gets a score of 2 on *X*, his or her predicted *Y* score will be 7. The appropriate regression approach would have been, however, to generate two separate lines because there are different intercepts for the two groups:

$$\text{Group 1 (Focal): } Y' = 5 + 2.0(X) \text{ and}$$
$$\text{Group 2 (Reference): } Y' = 1 + 2.0(X).$$

Thus, if an individual in Group 1 has a score of 2 on *X*, the predicted score would be 9 (2 more than would have been the case if one regression equation had been used). For an individual in Group 2 with a score of 2 on *X*, the predicted score would be 5 (2 less than would have been the case if one regression equation had been used).

It can also be the case that there are no differences in slopes or intercepts, but there are differences where the group scores fall on the same line (see Figure 9.6). In the regression sense, the relationship between *X* and *Y* for the two groups would not be considered unfair. Some do argue that this is unfair if the focal group scores lower than the reference group on both the test (*X*) and on the criterion (*Y*). For example, individuals from lower socioeconomic status

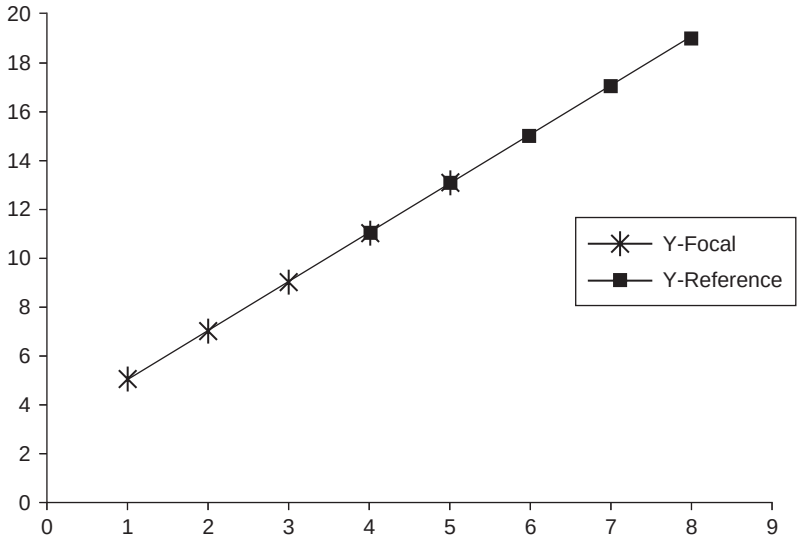


**Figure 9.6** A Single Regression Line for Both Focal and Reference Groups

families usually do more poorly on achievement tests than do those from higher socioeconomic status families. In addition, individuals from lower socioeconomic status families may also be more likely to be rated lower on a criterion measure (such as job performance). Those who claim that the process is unfair base their argument on systematic bias in both the test and criterion. However, this is an argument that has to be made on social or political grounds, not statistical grounds.

In general, tests are better if they do not result in differential validity coefficients for different identifiable subgroups. Usually, researchers and practitioners are interested in generalizing their knowledge claims about test scores as widely as possible.

## Extending the Inferences of Criterion-Related Validity Studies

For many small organizations, it is extremely difficult to conduct local validation studies. Assume, for example, that a company with 50 employees would like to institute a test as part of the selection system for new employees. However, gathering the needed data for a reasonable validity study to be carried out might take years before the sample size of new employees brought in is large enough to make any statistical inferences about the use of the test.

*Validity Generalization.* Many organizations face similar situations, where accountability for decisions must be demonstrated but practical constraints do not allow for appropriate statistical inferences to be generated. Some examples of these organizations would include (a) small firms wanting to use a test for employment selection purposes, (b) small social agencies interested in demonstrating that a particular program is more effective than another, and (c) small treatment centers wanting to demonstrate the use of a test in selecting appropriate treatment.

In such cases, the validity coefficient findings of previous studies are heavily relied on and generalized to the local context. The appropriateness of reliance on these findings, however, is based on the extent to which the local context (sample and criterion) can be assumed (and perhaps demonstrated to others) to be similar to the context in which the previous research was conducted. As Ghiselli pointed out some years ago (1959), validity coefficients for the same job in different contexts often give highly disparate validity coefficients.

*Meta-Analysis.* A meta-analysis attempts to generalize these disparate validity coefficients by summarizing a number of them across many contexts into a single index. Schmidt and Hunter (1977, 1990) popularized meta-analysis, where the results of many criterion-related studies using a predictor and criterion are pooled. Although it is beyond the scope of this book to describe in detail the meta-analysis processes, it is helpful to review the general process and understand the logic involved.

1. Calculate the desired descriptive statistic for each study available and average that statistic across studies.
2. Calculate the variance of the statistic across studies.
3. Correct the variance by subtracting out the amount due to sampling error.
4. Correct the mean and variance for study artifacts other than sampling error.
5. Compare the corrected standard deviation to the mean to assess the size of the potential variation in results across studies in qualitative terms. If the mean is more than two standard deviations larger than zero, then it is reasonable to conclude that the relationship considered is always positive. (Hunter & Schmidt, 1990, p. 82)

Thus, the first step is to identify all of the relevant published or unpublished validity coefficients. Then these are corrected for a number of artifacts. Schmidt and Hunter (1990) report no fewer than 11 sources of artifactual error in these coefficients that they recommend be corrected. They include sample size, unreliability of the predictor and criterion, range restriction, attrition, invalid measures of the predictor and criterion, transcription errors, and variance due to extraneous variables that are not controlled for. After making some or all of these corrections to the validity coefficients and then combining them (weighting some coefficients more than others), a single true validity coefficient emerges. This coefficient, then, represents the true magnitude of the relationship between predictor and criterion. The inference that can be drawn is that the predictor is related to the criterion at a certain level and is transportable across situations *(validity generalization)*. By using validity generalization (i.e., quoting a meta-analytically derived validity coefficient), one seemingly does not need to carry out an expensive, time-consuming local validation study in a specific context.

A critical feature of the meta-analytic process is in finding the results to use as fodder for the meta-analysis. The predictor has to be identified (e.g., measure of leadership style) as well as the criterion (e.g., task performance). An extensive search of both published and unpublished results is expected. The literature review is usually constrained over a certain time period (say the last 10 or 20 years). In addition, more constraints can be placed on the literature search that will make the results more specific to a particular context (such as the workplace).

As has been noted, the purpose of a meta-analysis is to generate a single corrected mean effect size that characterizes the relationship between two variables. For example, there may be an effect size in the population between intelligence and salary of, let's say, 0.25. This mean effect size has a standard deviation around it. A credibility interval can be set around the mean effect size using the corrected standard deviation. A credibility interval around the mean effect size that is large or contains zero suggests that the mean effect size (in our example, 0.25) is actually the mean of more than one subpopulation. For example, the relationship between intelligence and salary may be different for males compared to females (e.g., for

males it may be 0.35 and for females 0.00). A credibility interval that is small or does not contain zero suggests that the mean effect size is the mean of one population. Thus, credibility intervals provide information about whether moderator effects (such as gender in our example) need to be taken into account when interpreting the results.

The mean effect size also has a standard error around it reflecting sampling error. The standard error is used to set confidence intervals around the effect size. As in prior discussions about confidence intervals, these assist in determining the accuracy of the corrected mean effect size. Confidence intervals that do not include zero indicate the effect is significantly different from zero. Whitener (1990) provides a cogent and detailed review of credibility and confidence intervals.

While meta-analytic results are popular in the published research these days, and have their usefulness, meta-analysis is not without detractors. Nunnally and Bernstein (1994) say,

> Meta-analysis is extremely useful in aggregating the well-done studies hampered by small sample size. . . . Consequently, it can be a useful tool to integrate the literature. . . . Meta-analysis is no substitute for careful evaluation of individual studies' procedures and results, and was never intended as a "meat grinder" to average out results of studies that vary in their quality of execution. (p. 101)

Algera, Jansen, Roe, and Vijn (1984) say,

> Our critical comments on the Schmidt-Hunter approach to validity generalization do not imply that their work is not relevant for the theory and practice of personnel selection. . . . However, the ways in which they have worked out their ideas on validity generalization show fundamental shortcomings, needing correction in the future. As a consequence, their conclusions should be considered as premature. (pp. 208–209)

Informed consumers of meta-analytic results and users of validity generalization should be clear that the final true validity coefficient is only as good as the quality and completeness of the data (i.e., results from other studies) that went into generating the coefficient. As has been demonstrated, some of the corrections carried out as routine in meta-analysis (e.g., correction for unreliability) can have dramatic effects on the magnitude of the validity coefficients. The appropriateness of the corrections and weighting scheme should be readily apparent. Simply quoting the findings of a meta-analysis does not absolve the test user of being diligent in assessing the utility of the test for any specific context.

*Synthetic Validity.* Synthetic validity had been proposed as a potential alternative to local validation studies many years ago (Balma, 1959; Lawshe, 1952). Synthetic

validity assumes that at least two tests, called a test battery, will be used as predictors. Basically, the process is as follows. First, an analysis of the criterion (e.g., job performance, scholastic performance, social adjustment, etc.) is carried out such that the elements that make up the criterion are clearly articulated. Second, a test battery is compiled where each test is selected because it is expected to correlate with one or more of the elements. Third, the synthesis of the test battery validity coefficients into a single overall coefficient provides a *synthetic validity coefficient.* There are various ways to generate the synthetic coefficient. In this chapter, a couple of them will be described.

A simple but concrete example will help put all of these points in perspective. Assume we are interested in predicting who will successfully pass a driver's training course. The task is broken down into its elements: A = driver knowledge, B = psychomotor coordination, C = visual acuity, D = conscientious personality. Now, four tests are selected to create a battery: 1 = a paper and pencil test of driver knowledge, 2 = a test of psychomotor coordination, 3 = a test of acuity, and 4 = a test of conscientious personality.

The next step requires judgments to be made to determine the combinatorial rules for these tests. One of these is the *J* coefficient (Primoff, 1957; 1959). Generally, this process requires either ratings by SMEs or some empirically derived values to estimate the linkages between (a) the criterion and elements ($R_{ye}$) and (b) the tests and the elements ($B_{xe}$). These are then multiplied together over the elements and summed.

Continuing with the driving example, the $R_{ye}$ values were generated by having a group of 10 driver trainers weight each element in terms of its contribution to the overall criterion. They did this task under the restriction that the total value must add to 1.0. Their weightings were averaged, providing the $R_{ye}$ values.

Next, the $B_{xe}$ values have to be generated using one of a number of different ways. The literature can be searched and the average validity coefficient in previous studies between each test and the associated element found. Alternatively, concurrent criterion-related validity analyses with a sample of current successful and unsuccessful driver training students, correlating test scores with element scores can be carried out. Another way to generate the $B_{xe}$ values would be to have SMEs (e.g., driver trainers) estimate the validity coefficients given their knowledge of the test and the elements. Other techniques are also possible. However, regardless of which is used, the generated values now become our $B_{xe}$ values. The $R_{ye}$ and $B_{xe}$ values for each of the four elements for successful driving are as reported in Table 9.17.

Next, the $R_{ye}$ and $B_{xe}$ values for each element are multiplied and summed, providing the synthetic validity coefficient. In this example, the resulting coefficient of 0.35 leads to the conclusion that the test battery will predict driver success at a validity coefficient level of 0.35.

Hollenbeck and Whitener (1988) put a bit of a twist on the usual *J* coefficient procedure. They assessed several (13) jobs in terms of their elements. Although, across all jobs, 14 elements were identified, each job had only a few elements associated with it. Each element was rated as irrelevant, minor, or major for each job

**Table 9.17** The $R_{ye}$ and $B_{xe}$ Values to Generate the *J* Coefficient in Synthetic Validity

| *Element* | $R_{ye}$ | $B_{xe}$ | $R_{ye} \times B_{xe}$ |
|---|---|---|---|
| A (driver knowledge) | 0.50 | 0.30 | 0.15 |
| B (psychomotor coordination) | 0.20 | 0.25 | 0.05 |
| C (visual acuity) | 0.15 | 0.80 | 0.12 |
| D (conscientious personality) | 0.15 | 0.20 | 0.03 |
| | | | $\Sigma R_{ye} \times B_{xe} = 0.35$ |

by SMEs, where the values for each job were to total 1.0. Thus, if a job had two elements, one major and one minor, the weightings for that job would be 0.667 and 0.333 respectively. If a job had five elements, all rated as major, then they would be equally weighted at 0.20 each. These were called *element importance indices.*

Then employees in each job were given the tests associated with the elements that had been identified as relevant to their jobs. Next, supervisors rated each employee's performance on the elements. Individual test scores and individual ratings of performance were then standardized. Next, they created two matrices, with a row devoted to each employee. In the test matrix, each employee's standardized test value was multiplied by each element importance index. These were then summed to provide what will be called a test value. In the criterion matrix, each employee's standardized performance rating value was multiplied by each element importance index. These were then summed to provide what will be called a criterion value.

The synthetic validity index was calculated by correlating the test values with the criterion values across all 83 employees. It was 0.22, which is statistically significant based on a sample size of 83. This approach used synthetic validity for the entire population of employees in the firm. Because of the small number of employees in each job, it would have been impossible to carry out a traditional criterion-related validity study.

Synthetic validity is a very content-oriented approach to validity assessment and relies heavily on SMEs' opinions. The integrity of the synthetic validity approach hangs very much on how accurately the elements have been identified. If important ones are missed, then there will be a serious shortcoming in the synthetic validity coefficient. The links between element-criterion and test-element must be demonstrated to be reasonably and reliably representative of the true values for the synthetic validity coefficient to be believable (see Mossholder & Arvey, 1984, for a review of the various techniques). Synthetic validity has been used primarily in the context of employee selection testing and job performance. Given the move away from small job elements and toward broader job competencies in the job analysis literature, synthetic validity may

play a more prominent role than in the past. However, broadening its use with examples using predictors and criteria from other domain areas, such as scholastic performance or mental health, will be needed to generalize the use of this technique.

## Summary

The chapter was devoted to some of the methods used to assess the degree to which the inferences made from test scores are valid. Specifically, those methods that used the subjective assessment of item content and the relationships of the test scores to a criterion were reviewed. Many validation techniques were covered, including using

a. information from test-takers,
b. information from SMEs,
c. correlation and regression,
d. convergence and divergence, and
e. discriminant function analysis.

Other topics covered were

a. upper bounds of validity,
b. dealing with multiple criteria,
c. the standard error of measurement,
d. validity generalization, and
e. synthetic validity.

In the next chapter, the validity of inferences made from test scores will continue to be discussed. However, the primary focus will be on the internal structure of the test items.

## Problems and Exercises

1. Why is it important to ask test takers to provide feedback on test items?
2. Why is it important to ask SMEs to provide feedback on test items?

3. Suppose I have a Primary Grade Adjustment Test that I am interested in using to predict social adjustment in primary grade school (PGAT). Teacher assessments to social adjustment for a group of 50 students is regressed on age first (obtained $R^2$ value of 0.15,) and then on test score (obtained $R^2$ value of 0.40). What would the conclusions be?

4. What is the difference between predictive and concurrent criterion-related validity studies?

5. What are convergent and divergent assessment useful for?

6. What is the upper bound on my validity coefficient for a test with a reliability of 0.70 and a criterion with a reliability of 0.80?

7. What would be the corrected validity coefficient (uncorrected is 0.20) if the unreliability of the test and the criterion from the previous exercise (Item 6) is corrected?

8. Assume the correlation between GRE score and performance in first year of graduate school (measured via GPA) is calculated to be 0.15. Further assume that the range on GRE for the sample is restricted (standard deviation of the sample is 25 and standard deviation in the population is 100). What would be the corrected validity coefficient?

9. You are interested in seeing if a new test will help to determine who will be successful in a training program. The current rate of success is 70% for 200 students. When the new test is implemented, the rate increases to 85% for the 50 students who pass the test and are allowed into the program. Is this difference significant?

10. Assume that the base rate for success in a clinical psychology graduate program is 0.60. We want to improve this by using a test that is known to correlate with the criterion of graduate school success at 0.30. Then, in the next year, 30% of the applicants are allowed into the program. What is the proportion of students that will be successful given the new testing procedure?

11. Suppose that I want to use a test to help predict who will quit their jobs before the end of the three-month probation period. One hundred new applicants are hired and the test is administered to all of them the first day. Half of them quit before the end of the 3-month probation period. Discriminant function analysis is used to determine if the test is useful. The chi-square is significant and the success rate of predicting the "quitters" versus "stayers" is 80%. What would be the conclusion?

12. If the slopes of two regression lines are significantly different for two identifiable subgroups, what would be concluded?

13. Two identifiable subgroups have regression lines between the predictor and criterion that have equal slopes but different intercepts. Group 1 has an intercept of 5 and Group 2 has an intercept of 3. Which group will be unfairly treated?

14. What is the reason for wanting to use validity generalization?

15. What is meta-analysis?

16. Why is synthetic validity so dependent on the expertise of SMEs?