Imagine that you have applied for your dream job. You have spent time, energy, and perhaps money preparing yourself to be competitive for the position. After all of your preparation, you finally have the opportunity to apply for the job for which you have been working so hard. As part of the hiring process, the company requires that you complete a personality inventory. A week or two after completing the personality inventory and submitting your application, the company informs you that you are not going to be hired. Although they do not say so, you suspect that their decision is partially based on the “results” of the personality inventory. Aside from disappointment and perhaps anger, what kind of reactions would you have?

You would likely have several questions. You might wonder what exactly the personality inventory was supposed to measure. Is there any evidence that the inventory is in fact a good measure of whatever it is supposed to measure? Is there any logical or theoretical reason to believe that scores on the inventory are related to performance in the job that you wanted? Perhaps more importantly, are there any hard data showing that scores on the inventory are actually related to performance in your job?

In response to such questions, the human resources director of the company might suggest that the personality inventory is a worthwhile part of the hiring process. She might state that the company has been using it for years. In addition, she might assure you that, in her experience, the questionnaire is quite accurate and that it is useful for predicting who will be good employees. However, if she is going to be using the inventory to make such important
decisions, then she needs to have stronger evidence than “her experience” testifying to the accuracy and utility of the questionnaire.

Your questions about the personality inventory are questions of validity, which is perhaps the most important issue in psychological measurement. In this chapter, we define validity, we discuss its meaning and implications, and we discuss the kinds of evidence that are necessary for establishing validity in testing. In addition, we describe differences among several perspectives on test validity, we contrast validity with reliability, and we conclude by discussing the importance of validity in testing. As you will see, a test-user’s personal experience is inadequate as evidence for the test’s validity and its use.

What is Validity?

The concept of validity has evolved over more than 60 years, and various definitions of validity have been proposed. A rather basic definition of validity is “the degree to which a test measures what it is supposed to measure.” Although this definition is relatively common and straightforward, it oversimplifies the issue a bit. A better definition, reflecting the most contemporary perspective, is that validity is “the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses” of a test (AERA, APA, & NCME, 1999, p. 9). This more sophisticated definition has a number of important implications.

First, a measure itself is neither valid nor invalid; rather, the issue of validity concerns the interpretations and uses of a measure’s scores. Consider the Conscientiousness factor on the Revised NEO Personality Inventory (NEO-PI-R, Costa & McCrae, 1992). The NEO-PI-R is a multi-dimensional personality inventory, providing scores on five relatively independent domains, each of which includes six more narrow “facet” dimensions. One of the factors, or domains, included in the NRO-PI-R is labeled Conscientiousness. The Conscientiousness scale
includes 48 items, with each item presenting a statement regarding beliefs, interests, behaviors, and so on. The test authors offer a clear interpretation of the scores derived from the items on the Conscientiousness factor. According to the authors of the NEO-PI-R, high scores on this set of items reflect the tendency toward an “active process of planning, organizing, and carrying out tasks,” and they state that people with high scores on this set of items are “purposeful, strong willed, and determined” (Costa & McCrae, 1992, p. 16).

In terms of validity, the set of items themselves are neither valid nor invalid. Similarly, the scores derived from the 48 items are neither valid nor invalid. However, the authors’ interpretations of the scores might be valid or invalid. Are the authors correct in interpreting scores on the set of 48 items in terms of planfulness, organization, and determination?

Extending this notion beyond the interpretation of the scores, validity is related to the “proposed uses” of the scores. The NEO-PI-R Conscientiousness scale might be used by employers to screen applicants. Experts in human resources might believe that people who have relatively high scores on the Conscientiousness scale will be responsible, hard-working, motivated, and dependable employees. Based on this interpretation of scale scores, employers might use the Conscientiousness scale to identify people with a high level of Conscientiousness and to make hiring decisions based on this information. But is there good reason to believe that scores on the Conscientiousness scale really do provide information that differentiates potentially better and worse employees? That is, scores on the NEO-PI-R Conscientiousness scale might be interpreted validly as reflecting Conscientiousness, but are they truly predictive of the quality of an employee’s future performance?

Psychological measures are like hammers. Someone might tell you that a hammer is a useful tool, but the usefulness of a hammer actually depends on the job to be done. If you need
to drive a nail into a surface or if you need to remove a nail from a surface, then a hammer is enormously useful. If you need to hold down a piece of paper while you are working or if you need to break through a piece of sheetrock in a wall, then a hammer might indeed be useful. However, if you need to tighten a screw, saw a piece of wood, change a light bulb, or call a contractor to fix the hole in your wall, then a hammer is completely useless. So, it is somewhat simplistic and inaccurate to say that a hammer is a useful tool without regard to the way in which it will be used. Similarly, it is somewhat simplistic and inaccurate to say that a particular psychological measure, such as the NEO-PI-R Conscientiousness scale, is valid without regard to the way in which it will be interpreted and used. The scale’s scores might be interpreted validly as indicators of Conscientiousness, and they might help you select a conscientious contractor, but the scale is not valid as a measure of Intelligence or Extraversion.

Despite our insistence that validity is really about test interpretation and use (not about the test itself), test users often refer to the “validity of a test.” For example, you might hear someone state that “the Conscientiousness scale of the NEO-PI-R is valid.” There are at least two possible reasons that a test user might make a statement that seems to contradict our definition of validity. First, the test user might not have a sophisticated understanding of validity. Although many copyrighted psychological tests are marketed only to qualified professionals, not every professional has a deep understanding of the concept of validity. The second reason that you might hear such a statement is that it is simply a short-cut. That is, instead of saying “the Conscientiousness scale of the NEO-PI-R is valid as a measure of Conscientiousness,” we sometimes get a bit lazy and simply state “the Conscientiousness scale of the NEO-PI-R is valid.” Please do not let this confuse you. In measurement, validity is a property of the interpretations and uses of a test, it is not a property of the test itself.
A second important implication of the definition of validity is that validity is a matter of degree, it is not an “all-or-none” issue. That is, the validity of a test interpretation should be conceived in terms of strong versus weak instead of simply valid or invalid. There is no magical threshold beyond which validity is established. For test users, validity should be a deciding factor in their choice of psychological tests. Although such choices are based on a number of practical, theoretical, and psychometric factors, a test should be selected only if there is strong enough evidence supporting the intended interpretation and use. Alternatively, test users might need to choose among a set of possible tests, and they must weigh the relative strengths of evidential bases for the tests being considered. For example, there are a variety of scales that an employer might use to measure dependability, responsibility, motivation, and reliability of job applicants. The NEO-PI-R Conscientiousness scale is a reasonable choice, but employers should consider the relative strengths of alternatives that might be even better measures of the specific characteristics that they wish to assess.

A third important facet of validity is that the validity of a test’s interpretation is based on evidence and theory. In the introduction to this chapter, the hypothetical human resources director stated that, in her experience, the psychological tests were useful. This is not good enough. For a test user to be confident in an interpretation and use of test scores, there must be empirical evidence supporting the interpretation and use. In addition, contemporary views on validity emphasize the importance of grounding the interpretation and use of a test in a defensible psychological theory.

Although many well-developed psychological measures have strong evidence regarding the validity of their typical interpretations, many supposed psychological measures do not. For example, handwriting analysis is a popular method for “assessing” personality. Despite the
popularity and historical tradition of handwriting analysis, there appears to be little peer-reviewed scientific evidence that handwriting reveals anything about personality. Currently, many supposed psychological measures can be found on the internet. One example is variously known as “Color Quiz,” “Colorgenics” or the “Personality Color Test.” These tests are ostensibly based on “color psychology” as developed by Max Luscher (Luscher & Scott, 1969). When you take the “Color Quiz,” you are presented with eight colors and you are asked to select the colors in the order of your preferences (see http://www.colorquiz.com/). After completing this procedure twice, you receive a series of interpretations regarding your “sources of stress,” “restrained characteristics,” “desired objectives,” and your “actual problems.” The notion that your color preferences reveal something about your personality is an interesting idea, but is there any validity to this interpretation of color preferences? A quick survey of the scientific literature provides almost no support for the validity of color preferences as a measure of personality characteristics (e.g., Picco & Dzindolet, 1994).

Assuming that there is indeed little scientific support for the validity of color preferences as a measure of personality, it is interesting to examine the “evidence” presented on the Color quiz web site (see http://www.colorquiz.com/about.html). The web site poses the question to interested readers, “Is the test reliable?” We suspect that the authors of the web site are not using the term “reliable” in the true psychometric sense outlined in previous chapters. Instead, we suspect that the authors are intending to pose the question of validity – is the test meaningful and useful as a measure of personality? Given the apparent lack of scientific evidence for color preference as a valid measure of personality, you might not be surprised by the answer provided on the web site. Regarding the quality of the “Color Quiz,” the authors state, “We leave that to your opinion. We can only say that there are a number of corporations and colleges that use the
Lûscher test as part of their hiring/admissions processes.” Clearly, the web site implies that the Color Quiz is a valid measure of some aspects of personality and that it is in fact used to make real decisions. However, we suggest that a human resources director using any version of the “Color quiz” should be prepared to defend his or her hiring decisions in court. If the scientific evidence for the validity of “color tests” as measures of personality is as thin as it appears to be, then an applicant who is denied employment because of such a test would have legitimate reason to be angry and litigious.

The contemporary perspective on validity states that there must be psychological theory and empirical evidence supporting a particular interpretation of test scores. For example, are there strong data demonstrating that people who score relatively highly on the NEO-PI-R Conscientiousness scale are actually higher in “Conscientiousness” than those who score relatively lower? Is there evidence that students who perform well on the SAT actually obtain higher grades in college than do students who perform less well on the SAT? Is there anything beyond the assertion that “there are a number of corporations and colleges that use the Lûscher test as part of their hiring/admissions processes” to support the notion that color preferences actually reflect anything about an individual’s personality? Although Lûscher might offer theoretical reasons to suspect that color is somehow related to personality, such theory is not enough to argue that the Color quiz is valid as a measure of personality. For users to have confidence in the validity of test interpretations, there must be good empirical evidence supporting the interpretations. There must be data obtained from high-quality research, and these data must provide evidence for particular interpretations of test scores.

In the sections that follow, we will examine the kinds of scientific evidence that can be used to support the validity of test interpretations. As mentioned earlier, the concept of validity
has evolved over the years. For many years, the fields of psychology and education have seen validity as a three-faceted concept. From this traditional perspective, there are three types of validity – content validity, criterion validity, and construct validity. Although we will describe these concepts, we will emphasize a more contemporary perspective that highlights construct validity as the essential concept in validity (Messick, 1993). **Construct validity** refers to the degree to which test scores can be interpreted as reflecting a particular psychological construct.

In 1999, three major organizations in psychology and education outlined the contemporary perspectives on testing. The American Education Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) published a revision of the *Standards for Educational and Psychological Testing*. This publication outlined five types of evidence relevant for establishing the validity of test interpretations (AERA, APA, & NCME, 1999). The overall construct validity of test score interpretation depends on the content of the test content, the internal structure of the test, the psychological process used in test responses, the association among test scores and other variables, and the consequences of test use.

**Validity Evidence: Test Content**

One type of validity evidence relates to the match between the actual content of a test and the content that *should be* included in the test. If a test is to be interpreted as a measure of a particular construct, then the content of the test should reflect the important facets of the construct. The supposed psychological nature of the construct should dictate the appropriate content of the test. Validity evidence of this type is sometimes referred to as **content validity**, but there are two ways that content validity might be compromised.
One threat to content validity occurs when a test includes **construct-irrelevant content**. A test should include no content (e.g., items or questions) that is irrelevant to the construct for which the test is to be interpreted. Imagine that you are asked to develop a midterm test for a class in personality psychology, and the test is intended to measure “knowledge of Freud” as covered in the class lectures, discussion, and readings. In the class, the broad topics were covered – the structure of personality (i.e., id, ego, and superego), the stages of personality development, and defense mechanisms. Ideally, the content of the midterm test should include items representing a fair sample of these three topics, no more and no less. For example, biographical questions about Freud’s life should not be included on the test, because they were not covered in class and thus they are irrelevant to the construct of “knowledge of Freud as covered in the class lectures, discussion, and readings.” Test content that reflects issues, characteristics, or concepts that are irrelevant to the construct for which a test is to be interpreted is referred to as construct irrelevant content. Such content is actually extraneous to the core construct for which a test is to be interpreted, and its inclusion would reduce validity.

A second threat to content validity is **construct under-representation**. Although a test should not include content that is beyond its core construct, it *should* include the full range of content that *is* relevant to the construct, as much as possible. Again, a test intended to assess “knowledge of Freud as covered in the class lectures, discussion, and readings” should include content relevant to all three topics that were covered in class. A test that included content only relevant to personality structure and personality development would have weak validity measure of “knowledge of Freud as covered in the class lectures, discussion, and readings” because it fails to cover the content related to defense mechanisms. Such a test would suffer from construct under-representation, meaning that its actual content fails to represent the full scope of the
content implied by the construct that it is intended to measure. In sum, a test’s content should reflect the full range of the core construct, no more and no less.

Realistically, test developers and test users face a trade-off between the ideal of content validity and the reality of the testing situation. Earlier, we suggested that a test should include items that represent a fair sample of the construct-relevant content, no more and no less. However, there is no clear rule for what constitutes a “fair sample.” For practical reasons, a test developer might not be able to include content covering every facet or nuance of the construct to an equally thorough degree. For example, the instructor developing a test to assess “knowledge of Freud as covered in the class lectures, discussion, and readings” must consider the fact that students might have only 50 minutes to complete the test. Therefore, he or she might include questions regarding details of only some of the total content. For example, he or she might include questions on only three stages of Freud’s theory of personality development. So, the test might not cover every conceivable facet of the construct, but hopefully the selected items reflect a fair range of elements relevant to the construct. In sum, practical issues such as time, respondent fatigue, respondent attention, and so on place constraints on the amount of content that can be included in a measure.

**Face validity** is closely related to content validity. Face validity is the degree to which a measure appears to be related to a specific construct, in the judgment of non-experts such as test-takers and representatives of the legal system. That is, a test has face validity if its content simply looks relevant to the person taking the test. Face validity is not usually considered an important psychometric facet of validity – non-experts’ opinions have no direct bearing on the empirical and theoretical quality of a test.
Although face validity might not be crucial for test validity from a psychometric perspective, it might have important implications for its use. The apparent meaning and relevance of a test’s content might influence test-takers’ motivation to respond in a serious and honest manner. For example, consider a psychological inventory given to job applicants as part of the hiring process for a law enforcement agency. Applicants might assume that such a measure should include questions about problem solving, social skill, dependability, work ethics, and so on. If the inventory actually included questions about sexual attitudes or family history, then the job applicants might question the legitimacy or relevance of the entire testing procedure. Consequently, many applicants might respond randomly, respond in a way that presents a falsely positive image of themselves, or even refuse to complete the measure altogether. The utility of such a test would be almost entirely compromised. Therefore, a test with high face validity might be much more well-received by test-takers and potential test-users.

The difference between content validity and face validity is an important one. Content validity is the degree to which the content of a measure truly reflects the full domain of the construct for which it is being used, no more and no less. In a sense, content validity can be evaluated only by those who have a deep understanding of the construct in question. Experts in a field are the people who are in the best position to evaluate accurately the quality of a test of a construct within that field. Face validity is the degree to which non-experts perceive a test to be relevant for whatever they believe it is being used to measure. Although test-takers’ beliefs about a test might affect their motivation and honesty in responding to a test, test takers are not often experts on the theoretical and empirical meaning of the psychological constructs being assessed by the tests. Thus content validity, but not face validity, is an important form of evidence in the overall evaluation of construct validity.
Validity Evidence: Internal Structure of the Test

A second issue related to the validity of a test interpretation concerns the internal structure of a test. A test’s internal structure is the way that the parts of a test are related to each other. For example, some tests include items that are highly correlated with each other, but other tests include items that fall into two or more clusters. As we will discuss, the conceptual basis of a construct has implications for the internal structure of a measure of the construct. Therefore, an important validity issue is the match between the actual internal structure of a test and the structure that the test should possess. For a test to be validly interpreted as a measure of a particular construct, the actual structure of the test should match the theoretically-based structure of the construct.

For example, we might wish to evaluate measures of self-esteem. The Rosenberg self-esteem inventory (RSEI; Rosenberg,) is perhaps the most commonly-used measure of self-esteem in psychological research. The RSEI is often used as a measure of a single coherent construct – global self-esteem. Global self-esteem is one’s overall evaluation of one’s self-worth, and the RSEI includes ten items such as “I take a positive attitude toward myself” and “At times I think I am no good at all” (note that this item is reverse-keyed). Test users who intend to interpret scores on the RSEI as a measure of global self-esteem should expect to find a particular structure among the ten items. Specifically, if test users theorize that global self-esteem is indeed a single coherent construct and they believe that the RSEI is indeed valid as a measure of global self-esteem, then they should find that all of the items on the RSEI are highly correlated with each other, forming a single tight cluster of items. That is, if the RSEI is indeed valid as a measure of global self-esteem, then responses to the test items should exhibit a specific structure that is consistent with a conceptual definition of the construct.
However, our expectations might be quite different for another measure of self-esteem. The Multidimensional Self-esteem Inventory (MSEI; O’Brien & Epstein, 1988) was designed to measure global self-esteem along with eight components of self-esteem. The test authors state that the conceptual model underlying the MSEI “specifies two primary levels within the hierarchy of self-esteem. The first level corresponds to global self-esteem. This level is concerned with the person’s most basic, widely generalized evaluative feelings about him/herself. The second level corresponds to self-evaluations at an intermediate level of generality, which are referred to as components of self-esteem” (p.7). This conceptual perspective on self-esteem was based on previous research suggesting that the components of Competence, Likability, Lovability, Personal Power, Moral Self-approval, Body Appearance, and Body Functioning capture many of the events that affect self-esteem. Thus, the authors argue that these components reflect most of the experiences that typically affect self-esteem.

If the MSEI is indeed valid as a measure of these components of self-esteem, then responses to the test items should exhibit a specific structure that is consistent with the multidimensional conceptual definition of the construct. That is, the items on the MSEI should form separate clusters, they should not form one large cluster. In fact, the items should more-or-less form one cluster for each of the components.

To evaluate the internal structure of a test such as the MSEI, test-developers often use a statistical procedure called factor analysis. Some items on a test might be more strongly correlated with each other than with other items, and items that are highly correlated with each other form clusters of items, called “factors.” Factor analysis is a tool that can help us identify the presence and nature of “factors” existing within a set of items.
Consider a six-item personality questionnaire. Say the questionnaire includes the following six adjectives describing various personality characteristics: talkative, assertive, imaginative, creative, outgoing, and intellectual. For each item, respondents rate themselves on a 1 to 5 scale, in terms of how much the item describes themselves. There are six items, but how many different aspects of personality are really being measured by this short questionnaire? Do the items reflect six completely different and unrelated aspects of personality, or do they reflect only one or two core dimensions of personality?

Take a moment to think about the six items and group them into clusters that share some common meaning. In other words, group them into clusters of items that more or less seem to mean the same thing. Some of us might suggest that the brief questionnaire really only includes two clusters of items. For example, some might suggest that talkative, assertive, and outgoing are three variations on one basic idea (let’s call it Extraversion) and that imaginative, creative, and intellectual are three variations on another basic idea (let’s call it Openness to Experience). From this perspective, these six personality adjectives have a two-factor structure – there are two factors underlying these six items. Others might suggest that there are three clusters of items. Talkative, assertive, and outgoing might go together, and imaginative and creative might go together, but “intellectual” is something different from the other five items. From this perspective, the six items have a three-factor structure.

An important difficulty with this approach – an approach only based on our interpretations of the meaning of items – is that there is little way to evaluate which perspective is the best. That is, if you believe that there is a two-factor structure to the questionnaire, but I believe that there is a three-factor structure, then how could we determine who is correct, or if either one of us is correct?
Rather than relying on idiosyncratic interpretations of the meaning of items, test developers and users often prefer to base their arguments on hard data. So, we might give the six-item questionnaire to a sample of 100 respondents, enter their data, and compute the correlations among the six items. We would then use the correlations to help us identify and interpret the way in which the items cluster together.

For example, look at the correlation matrix presented in table XX-1. Notice that three of the items – talkative, assertive, and outgoing – are all strongly correlated with each other. An individual who rates herself as high on one of these three items is likely to rate herself as high on the other two items. We also see that the other three items – imaginative, creative, and intellectual – are strongly correlated with each other. An individual who rates himself as high on one of these items is likely to rate himself as high on the other two. Importantly, we also see that these two clusters of items are quite independent. Notice that the correlation between talkative and creative is zero, as is the correlation between talkative and imaginative, between outgoing and intellectual, and so on. That is, the fact that an individual rates herself as highly assertive, talkative, and outgoing tells us nothing about her level of creativity, imagination, or intellect.

From this pattern of correlations, we would likely conclude that the six-item personality questionnaire has a two-factor structure. Three items cluster together into one factor, the other three cluster into a second factor, and the factors are relatively independent.

By examining the pattern of correlations, we have performed a very basic factor analysis. Unfortunately, such a simplistic “eyeballing” approach rarely works with real data. In real data, there are often many more items to be considered. In the current example, we examined only six items, but many measures include considerably more than six items. For example, the Conscientiousness scale of the NEO-PI-R includes 48 items. Difficulty arises because a larger
number of items means a much larger number of correlations to examine. Indeed, even a few more items means many more correlations. In fact, if we were to examine a correlation matrix among 48 items, we would have to inspect more than 1,100 correlations! Obviously, eyeballing such a huge correlation matrix is a nearly impossible task. Aside from the sheer number of correlations in most real data, the pattern of correlations is almost never as clear as it appears to be in Table XX-1. The made-up correlations in Table XX-1 include a few very strong positive correlations and a few zero correlations, nothing in between. In real data, correlations often are closer to .18 or -.32 than .70. Therefore, the clusters of items in real data are not so readily apparent as they are in Table XX-1, and this makes the process of evaluating internal structure potentially very difficult.

Factor analysis is a statistical procedure that simplifies this process for us. Rather than needing to visually inspect a matrix of dozens or even hundreds of correlations, we use factor analysis to process the correlations for us. The results of a factor analysis allow us to address at least three inter-related issues related to the internal structure of a measure. First, factor analysis helps us understand the number of factors within a set of items. As described earlier, many social and personality psychologists would theorize that global self-esteem is a single coherent construct. Therefore, if the RSEI is indeed valid as a measure of global self-esteem, then responses to the ten items on the RSEI should form only a single factor. If analyses revealed that the RSEI items formed two or more factors, then we would begin to question the validity of the RSEI as a measure of global self-esteem. Thus, the number of factors is an important facet of evaluating the internal structure of a measure.

A second internal structure issue for which factor analysis is useful is determining which items are linked to which factors. In developing a measure, our conceptual understanding of a
construct might lead us to generate specific items to reflect particular aspects of the construct. Thus, to evaluate the quality of the measure, we would need to ensure that, in fact, the items that are intended to reflect a particular factor actually are connected to that factor and to no other factors. For example, the authors of the MSEI report the results of a factor analysis of all the items on their inventory (O’Brien & Epstein, 1988, pp. 14-15). They demonstrated that nearly every item was connected strongly to the component that it was written to reflect and weakly connected to all other components. For example, the ten items written for the Competence component were clearly connected to it and to no other. Similarly, the ten items written for the Moral Self-approval component were clearly connected to it and to no other. The results were not perfect though. For example, although the ten items written for the Body Appearance component were connected to it and no other, two “Global self-esteem items and three Likability items were also connected to the Body Appearance component. Despite a few imperfections in the internal structure of the MSEI, the authors seem generally satisfied that the factor analysis provides adequate support for the validity of the internal structure of the MSEI.

A third important use of factor analysis is to reveal the associations among several factors. If our theory of self-esteem suggests that all the components of self-esteem are independent, then we should find that the self-esteem scales are uncorrelated with each other. However, if our theory suggests that the components are associated with each other in a particular way, then a factor analysis should reveal that particular pattern of associations. Although the authors of the MSEI do not specify their expectations about the links among the components of self-esteem, they conducted a factor analysis to investigate those associations (O’Brien & Epstein, 1988, pp. 15-16). Their analysis revealed an interesting three-factor structure to the MSEI scales. Some MSEI scales (e.g., Global self-esteem, Competence,
Personal Power) clustered together to form an “Overall self-evaluation” factor, which the authors interpreted partly as reflecting the “ability to have an active and direct impact on the world by demonstrating capabilities, leadership abilities, body agility, and self-discipline” (p. 16). Other MSEI scales (e.g., Lovability, Likability) clustered together to form a “Social Self-esteem” factor, which the authors suggest “is dependent on the social feedback and approval or disapproval received from significant others” (p. 16). Two scales – Moral self-approval and a Defensive self-enhancement scale – clustered together to form a “Defensiveness and private self-evaluation” factor, which the authors suggest has “little in the way of objective or tangible social feedback” (p. 16).

In sum, the internal structure of a test is an important issue in construct validity. A test’s internal structure should correspond with the structure of the construct that the test is intended to measure. Typically, internal structure is examined through the correlations among the items in a test and among the subscales in a test (if there are any), and researchers often use factor analysis in this process.

Validity Evidence: Response Processes

A third type of validity evidence is the match between the psychological processes that respondents actually use when completing a measure and the processes that they should use. Many psychological measures are based on assumptions about the psychological processes that people use when completing the measure. For example, a researcher developing a measure of Extraversion might include an item such as “I often attend parties,” and this researcher might assume that respondents will read the item, search their memories for the number of times that they have attended parties, and then make a judgment about whether that number qualifies as
“often.” If participants do not use such a process, then the measure might not provide scores that are interpretable as the test developer intended.

A recent study of the effect of control deprivation on cognitive performance illustrates the sometimes subtle problem of response processes. Previous research has suggested that people who lack control over their outcomes in one task will show impaired performance on subsequent tasks. In the study, participants first engaged in a task in which some of them were able to exert control over a noise but others were not. In the “noise control” task, all participants were exposed to a loud buzzing noise, and they were instructed to learn a sequence of keystrokes that would temporarily suppress the noise. During this task, half of the participants were required to learn a very easy sequence, which ensured that they would eventually control the noise. The other half of the participants were required to learn an impossible sequence, which ensured that they would lack control over the noise. After the noise control task, all participants completed a series of word tasks. They were given a list of scrambled words (e.g., pynhm) and were required to identify the correct word (e.g., nymph). Participants were instructed to proceed one word at a time, and to proceed to the next word only after completing the prior word. The total number of words correctly unscrambled was taken as the measure of cognitive performance. The researchers hypothesized that control deprivation on the noise task would impair attention, which would produce lower scores on the word scramble task.

Take a moment to consider the psychological process that participants were assumed to use when responding to the scrambled words. The researchers implicitly assumed that participants would need to devote cognitive attentional resources to the word task. Although cognitive factors such as intelligence and previous experience with word puzzles could also affect performance on the word task, such factors were reasonably assumed to be constant across
noise control groups because participants had been randomly assigned to the groups. Thus, the researchers assumed that, if they found a group difference in the mean number of words correctly unscrambled, it would be because control deprivation would impair some participants’ ability to devote full attention to the word task. The impairment of cognitive resources would reduce those participants’ ability to concentrate on the word task, which would in turn decrease their performance on the task.

Results did not support the predictions – participants who had experienced control deprivation on the noise task were able to complete just as many words as were participants who had not experienced control deprivation. Many researchers would have taken this as evidence against the hypothesized effect of control deprivation on cognitive performance. However, the researchers who conducted this study paid careful attention to the participants’ responses to the measure of cognitive performance. When examining participants’ responses, the researchers realized that some participants had not followed the instructions to continue to the next scrambled words only after successfully unscrambling the previous word. That is, some participants either had forgotten the instructions or had consciously chosen to ignore them. Closer inspection revealed a group difference in this “cheating” behavior – more participants in the control deprivation group “cheated,” as compared to the control group.

What does this group difference suggest about the psychological processes that affect the measure of cognitive performance? Although the researchers had assumed that differences in performance would primarily reflect differences in attentional processes, their inspection revealed at least one other process that affected performance. Specifically, “adherence to instructions” also had an effect on performance because participants who “cheated” were able to unscramble more words correctly. It is possible that their hypothesis was actually correct – that
control deprivation impairs attention, which reduces cognitive performance – but participants who had experienced control deprivation also cheated which inflated their scores on the measure of cognitive performance.

Hopefully, this example illustrates the important point that construct validity can be evaluated in part by considering the processes involved in responding to a measure. In this example, the word task did not have strong validity as a measure of attention-based cognitive performance. The researchers’ laudable attention to their data revealed that the word task also might have been affected by participants’ adherence (or lack thereof) to the task instructions. In sum, their research is inconclusive regarding their original hypotheses, but their attention to response processes raises intriguing issues regarding the association between control deprivation and rule-adherence.

Validity Evidence: Associations with other Variables

A fourth type of validity evidence involves the associations between test scores and other variables. The contemporary view of validity emphasizes the theoretical understanding of the construct for which test scores are to be interpreted. If respondents’ test scores are to be interpreted as reflecting the respondents’ standing on a specific psychological construct, then our theoretical understanding of that construct should lead us to expect that test scores will have particular patterns of associations with other variables.

For example, if the RSEI is to be interpreted as a measure of global self-esteem, then we should think carefully about the nature of global self-esteem. Specifically, we should consider the way in which global self-esteem is associated with other psychological constructs such as happiness, depression, intelligence, social motivation, assertiveness, and so on. Our theoretical perspective on self-esteem might lead us to believe that people with high levels of self-esteem
should be relatively happy, relatively non-depressed, and relatively highly socially motivated. In addition, our theoretical perspective might state that self-esteem is unrelated to intelligence – people with low levels of self-esteem are equally intelligent, on average, as a people with high levels of self-esteem. Thus, our theoretical perspective on self-esteem suggests a particular pattern of associations between self-esteem and other psychological constructs. If RSEI scores can be validly interpreted as a measure of self-esteem (as we understand it), then we should find a particular pattern of associations between RSEI scores and measures of happiness, depression, social motivation, and intelligence.

Thus, the fourth type of validity evidence involves the match between a measure’s actual associations with other measures and the associations that the test should have with the other measures. If a test’s pattern of actual correlations with other tests matches the pattern of correlations that our theoretical perspective leads us to expect, then we gain evidence supporting the interpretation of the test as a measure of the construct in question. For example, imagine that we conduct a study in which respondents complete the RSEI along with measures of happiness, depression, social motivation, and intelligence. If we find that the RSEI is indeed positively correlated with happiness and social motivation, negatively correlated with depression, and uncorrelated with intelligence, then we gain confidence that RSEI scores can be interpreted validly as a measure of self-esteem. Conversely, if a test’s pattern of actual correlations with other tests does not match the pattern of correlations that our theoretical perspective leads us to expect, then we have obtained evidence against the interpretation of the test as a measure of the construct in question. If we find that RSEI scores are uncorrelated with happiness and social motivation, then we lose confidence that it should be interpreted as a measure of self-esteem.
When evaluating the pattern of correlations between a measure and other measures, it is important to consider convergent evidence. **Convergent evidence** is the degree to which test scores are correlated with tests of related constructs. In the global self-esteem example, our theoretical perspective states that happiness and social motivation are related to self-esteem. In addition, our theoretical perspective states that depression is related to self-esteem, albeit in a negative direction. Thus, if our research reveals that the RSEI is in fact positively correlated with measures of happiness and social motivation and that it is negatively correlated with measures of depression, then we have obtained convergent evidence.

Often in the process of evaluating the validity of test interpretations, researchers will ask respondents to complete several different measures of the same construct. For example, we might ask our respondents to complete the RSEI, along with other measures of self-esteem such as the Coopersmith self-esteem inventory and MSEI. We would naturally expect to find strong positive correlations between the RSEI and other measures of self-esteem. If we failed to find this, then we would question the validity of the RSEI as a measure of self-esteem. Similarly, researchers might include responses by “informants” to evaluate a test. For example, we might ask each of our respondents to recruit a close acquaintance, and we could ask the acquaintances to rate the self-esteem of the respondents. Although we might not expect to find an extremely high correlation, we would likely expect to find a positive correlation between our respondents’ self-reported RSEI scores and the RSEI ratings provided by their acquaintances. That is, we would expect to find that the respondents who described themselves as having relatively high self-esteem were described by their acquaintances as having relatively high self-esteem. In sum, convergent evidence often comes in the form of correlations among different ways of measuring the same construct.
When evaluating the pattern of correlations between a measure and other measures, we must also consider discriminant evidence. **Discriminant evidence** is the degree to which test scores are uncorrelated with tests of unrelated constructs. In the global self-esteem example, our theoretical perspective states that intelligence is unrelated to self-esteem. Thus, our research should reveal that the RSEI is in fact uncorrelated (or only weakly correlated) with measures of intelligence. If we found that RSEI scores were actually positively correlated with intelligence, then the RSEI would lack discriminant validity as a measure of self-esteem. That is, the RSEI would appear to measure more than just self-esteem.

Discriminant evidence is an important but perhaps subtle concept. Whether a measure is being used for research purposes or for applied purposes, test users must be confident that they know exactly which psychological variables are being measured. Consider a study that might be conducted by a developmental psychologist. The researcher might be interested in the association between self-esteem and academic ability. The researcher might recruit a sample of high school students to complete the RSEI, and he might also obtain students’ permission to get their standardized academic achievement test scores from their academic records. He computes the correlation between RSEI scores and academic achievement scores, and he finds a correlation of .40. He interprets this as indicating that students who have relatively high self-esteem tend to perform relatively well in school. Based on these results, he might even suggest that schools should devote resources toward increasing students’ self-esteem. The conclusions might be considered to have important implications for psychological theory. They might even influence the way that school systems spend money.

Before putting too much confidence in the researcher’s conclusions, we should carefully consider the methods used in his study, including the discriminant validity of his supposed
measure of self-esteem. Whether his conclusions are correct depends in part upon the degree to which the RSEI has discriminant validity as a measure of self-esteem. If scores on the RSEI are in fact highly correlated with measures of intelligence, then the RSEI lacks discriminant validity and is at least partially a measure of Intelligence. Thus, the correlation found by the researcher might be more accurately interpreted as indicating that students who have relatively high Intelligence tend to perform relatively well in school. Indeed, it is possible that the trait of self-esteem is actually unrelated to academic performance and that the researcher has made a serious error in his conclusions. This error could have harmed psychological theory, and it could lead to wasteful spending of limited school resources.

Another common distinction related to this type of evidence is the distinction between concurrent validity evidence and predictive validity evidence. Concurrent validity evidence refers to the degree to which test scores are correlated with other relevant variables that are measured at the same time as the primary test of interest. For example, the website of the College Board, the company that administers the SAT, tells students that the SAT Reasoning test “is a measure of the critical thinking skills you'll need for academic success in college” (College Board, 2006). How could we evaluate the validity of the SAT as a measure of skills needed for academic success? One possibility would be to have students complete the SAT during their senior year in high school, and then correlate their SAT scores with their high school Grade Point Average (GPA). That is, we could examine the correlation between SAT scores and GPA scores that are obtained at more-or-less the same time that students responded to the SAT.

Predictive validity evidence refers to the degree to which test scores are correlated with relevant variables that are measured at a future point in time. For example, another way of evaluating the validity of the SAT is to have students complete the SAT during their senior year in high school,
and then correlate their SAT scores with the GPA obtained during their freshman year in college. That is, we could examine the correlation between SAT scores and GPA scores that are obtained a year or more after students responded to the SAT. Although the distinction between concurrent and predictive validity evidence is traditionally important, the larger point is that both refer to the match between test scores and other relevant variables. As we discussed earlier, the match between test scores are other relevant variables is referred to as convergent validity. Therefore, concurrent validity or predictive validity are essentially varieties of convergent validity.

There are many important issues involved in evaluating convergent and discriminant evidence, and such evidence is arguably the most important facet of validity. The current description has focused mostly in their conceptual meaning and importance. Because of the importance of convergent and discriminant evidence, the next chapter will explore many additional issues in greater detail.

Validity Evidence: Consequences of Testing

As discussed earlier, one key difference between the contemporary perspective on validity and the traditional three-faceted perspective on validity is that the contemporary perspective emphasizes the primacy of construct validity over content validity and criterion validity (more will be said about this later as well). However, an even more radical and contentious difference may be the assertion that the social consequences of testing is a facet of validity.

The Standards for Educational and Psychological Testing states that validity includes “the intended and unintended consequences of test use” (AERA, APA, & NCME, 1999, p. 16). More specifically, Cronbach (1988) states that test developers, users, and evaluators “have an obligation to review whether a practice has appropriate consequences for individuals and
institutions and especially to guard against adverse consequences” (p. 6). For example, if a construct and its measurement seem to benefit males more than females in the workplace, then we should be concerned about the use of the test. Are test scores equally valid for males and females as a measure of the intended construct? How large is the difference in the benefits for males and females?

The suggestion that the consequences of testing are linked to construct validity has generated quite a bit of debate. Almost everyone would agree that test users, test takers, and policy makers should be concerned about the possibility that a testing program might unfairly and adversely affect some people more than others. However, not everyone agrees that the consequences of a testing program should be considered a facet of the scientific evaluation of the meaning of test scores. For example, Lees-Haley (1996) considers the assertion that validity includes an assessment of the actual and potential consequences of test use, and he poses the questions, “But whose consequences? And who will decide? Enemy Psychologists? Pacifists? Generals? Whose social values shall we use to assess the consequential validity of these tests?” (p. 982). Clearly, the infusion of value judgments into an ostensibly objective scientific process raises some interesting issues. Lees-Haley (1996) bluntly states that “Consequential validity is a dangerous intrusion of politics into science” (p. 982).

Proponents of consequential validity would respond by arguing that science can never be separated from personal and social values. The questions that scientists investigate are shaped partially by society’s values and by their own personal values. The theoretical assumptions that scientists make are partially shaped by value judgments, and even the labels that scientists attach to their theoretical concepts are partially shaped by values. As an example based on Messick’s (1993) important work, consider two psychologists who are developing a theory around a
personality construct. One psychologist believes that the construct should be called “flexibility versus rigidity” to differentiate people who can adapt their cognitive and behavioral tendencies in response to changing circumstances from people who tend to retain cognitive and behavioral stability. The other psychologist considers the construct and the psychological difference that it is intended to reflect, but she believes that the construct should be called “confusion versus consistency.” Which labels are “scientifically” correct? Should a high level of cognitive and behavioral variability be considered flexibility or should it be considered confusion? Should a low level of cognitive and behavioral variability be considered rigidity or should it be considered consistency?

Similarly, consider the following personality characteristic - the tendency to experience, recognize, monitor, and understand emotional reactions. Imagine that a test developer creates a measure of this characteristic and happens to find that females score higher than males, on average (social scientists seem to love searching for gender differences). What would the test developer choose to call the test and the construct that it is intended to measure? Knowing that females tend to score higher than males, would the test developer be likely to call it “emotional sensitivity”? If results had indicated that males tended to score higher than females, would the test developer instead choose to call it “emotional intelligence”? Furthermore, imagine that a human resources director was told that you have a high level of “emotional sensitivity.” Would his impression of you be different than if he had been told that you have a high level of “emotional intelligence”? Which label would you prefer? Would you have been hired if the human resources director believed that you were “intelligent” instead of “sensitive?”
The point here is that value judgments have potentially subtle (and sometimes not so subtle) influences on the scientific process. Proponents of consequential validity argue that such influences should be recognized and evaluated as clearly as possible in a testing context.

The issue of test bias will be discussed in greater detail in Chapter XX; however, some brief comments are relevant here. Earlier we suggested that consequential validity concerns the possibility that some people are unfairly and adversely affected by a testing program. It is important to recognize the difference between fairness in testing and the consequences of testing.

A test can have adverse consequences for a person or a group of people, and yet the test might still be fair. For example, imagine that females tend to score higher than males on measures of Conscientiousness. And imagine that a human resources director uses a Conscientiousness questionnaire in the hiring process, resulting in fewer males being hired than females. Is this fair? Does it constitute “adverse consequences” for males?

In this case, fairness depends on the nature of the gender difference? Why might females and males have different scores on the test? One possibility is that the test is biased. That is, the test does not measure the Conscientiousness equally well for all people – for whatever reason, it is a good measure of Conscientiousness for females but not for males. This is clearly an issue of validity – the test is not equally valid for all people. Therefore, hiring decisions made partially on the basis of test scores are unfairly biased against males.

A second possibility is that the test is not biased. That is, the test does in fact measure Conscientiousness equally well for females and males, it just so happens that females truly tend to have a higher levels of Conscientiousness than do males. In this case, hiring decisions made partially on the basis of test scores are not unfairly biased against males (assuming that there is empirical data demonstrating that Conscientiousness does predict job performance).
What about the consequences of the testing program? Whether or not the test is fair or biased against males, males are adversely affected by the test scores. However, the test is unfair only if the test is not equally valid for females and males. The existence of a group difference in test scores, by itself, does not tell us whether the test is fair or not, in terms of its validity as a measure of a psychological characteristic. In Chapter XX, we discuss these issues in more detail, and we present ways of evaluating whether a test is actually biased between groups.

Other Perspectives on Validity

So far, this chapter has conceptualized validity in terms of the degree to which test scores can be accurately interpreted as reflecting a particular psychological construct. This perspective assumes a link between test scores and a construct with a clear theoretical basis. Indeed, the types of evidence outlined above hinge on the fit between various aspects of test responses and various aspects of a construct’s theoretical basis. This perspective is, in essence, a theory-testing view of validity, but there are at least two alternative perspectives on validity. You might encounter these alternative perspectives in your readings or discussions, so we briefly describe them here.

**Criterion validity** is an alternative perspective that de-emphasizes the conceptual meaning or interpretation of test scores. Test users might simply wish to use a test to differentiate between groups of people or to make predictions about future outcomes. For example, a human resources director might need to use a test to help predict which applicants are most likely to perform well as employees. From a very practical standpoint, she might not care about the particular psychological construct that the test might be measuring, and she might not be concerned about the theoretical implications of high and low test scores. Instead, she focuses
on the test’s ability to differentiate good employees from poor employees. If the test does this well, then the test is “valid” enough for her purposes.

From the traditional three-faceted view of validity, criterion validity refers to the degree to which test scores can predict specific criterion variables. From this perspective, the key to validity is the empirical association between test scores and scores on the relevant criterion variable, such as “job performance.” Concurrent validity and predictive validity have traditionally been viewed as two types of criterion validity, because they refer to the association between test scores and specific criterion variables. According to the traditional perspective on criterion validity, the psychological meaning of test scores is relatively unimportant – all that matters is the test’s ability to differentiate groups or predict specific outcomes.

Although criterion validity is a relatively common term in psychometrics and has traditionally been viewed as a separate type of validity, the contemporary perspective suggests that evidence of criterion associations should be subsumed within the larger and more important concept of construct validity (Messick, 1993). From this perspective, criterion validity is not sufficient on its own, even for purely practical or applied contexts such as employee screening. Messick (1993) suggests that “even for purposes of applied decision making, reliance on criterion validity or content coverage is not enough. The meaning of the measure, and hence its construct validity, must always be pursued – not only to support test interpretation but also to justify test use” (p. 17).

Another alternative perspective on validity emphasizes the need to learn what test scores mean, rather than testing specific hypothesis about test scores. That is, instead of assuming that the theoretical basis of a construct is fully formed and then testing specific hypotheses regarding that theory, test developers and users can evaluate a test by assuming that the meaning of test
scores is itself an interesting and important question to be addressed. An “**inductive**” approach to validity proceeds by examining the associations between test scores and a large set of potentially important and relevant psychological variables (Gough, 1965; Ozer, 1989). The perspective emphasized in this chapter has been called a “deductive” approach to validity (Ozer, 1989), because test evaluation proceeds by deducing a particular hypothesis from the theoretical basis of a construct and then empirically evaluating the accuracy of the hypotheses. Whereas the deductive perspective is a theory-testing approach, the inductive perspective is a more exploratory approach. The goal of an inductive approach is to understand the full meaning of test scores, beyond the meaning that might be constrained by reference to a specific construct. From this approach, researchers “allow constructs to evolve and change as a planned part of the test construction process itself” (Tellegen & Waller, in press).

The inductive approach to validity might be most relevant within a research context, and it can be seen as a back-and-forth process. In an applied context, test developers and test users will probably focus on a test for the purposes of a well-specified use, such as predicting job performance. In a research context, test developers and test users might be interested in tackling a new area of interest and developing a theoretical foundation for the area. In such a case, test construction and evaluation goes hand-in-hand with the researcher’s evolving understanding of the constructs being measured. For example, Tellegen and Waller (in press) describe the development and evaluation of the Multidimensional Personality Questionnaire (MPQ). In its current version, the MPQ consists of 11 primary personality factor scales (e.g., Social Potency, Achievement, Stress Reaction), which are clustered into 4 broad traits (e.g., Positive Emotional Temperament, Negative Emotional Temperament). The development of the MPQ was motivated by “a desire to clarify and highlight the nature of several unmistakably important or focal
dimensions repeatedly emphasized or adumbrated in the personality literature” (p. 10). During the years-long development of the MPQ, items were written, administered, analyzed, and re-written repeatedly. During this process, the researchers refined their understanding of the constructs that seemed to be emerging from the MPQ.

Although the inductive approach to test validity can be informative in terms of expanding our understanding of a measure’s theoretical and practical implications, it is not commonly emphasized in the testing literature. Again, tests are typically developed with a focus on fairly specific constructs, and test developers usually spend their energy evaluating test score interpretation with sharp regard to those specific constructs. Less often do test developers spend time and effort examining a more comprehensive view of the test’s implications.

**Contrasting Reliability and Validity**

With the concept of validity now in mind, it might be useful to contrast validity and reliability. These two concepts are fundamental to a sophisticated understanding of psychometrics, and it is important to understand the difference clearly.

Recall from previous chapters that a test’s reliability is the degree to which differences in test scores reflect differences among people in their levels of the trait that affects test scores. At this point, we might add a bit to that definition, and suggest that a test’s reliability is the degree to which differences in test scores reflect differences among people in their levels of the trait that affects test scores, *whatever that trait might be*. That is, we can discuss the reliability of a particular test without even being aware of the potential interpretation of test scores or the nature of the trait being measured by the trait.

On the other hand, validity is intrinsically tied to the interpretation of test scores and the nature of the trait supposedly being assessed by the measure. In a sense, reliability might be
considered to be a property of test responses, whereas validity is a property of the interpretation of test scores. That is, reliability is a relatively simple quantitative property of test responses, but validity is an issue more tied to psychological theory and to the implications of test scores.

Even though they are separate concepts, reliability and validity are linked both conceptually and statistically. Conceptually, for many areas of interest in the behavioral sciences, validity requires reliability. For example, Intelligence is usually conceptualized as a psychological trait that is quite stable across time and situations – your true level of Intelligence does not change very much from week to week or month to month. Therefore, a test that is intended to be a measure of Intelligence should result in scores that are reasonably stable across time. That is, a valid test of Intelligence will be reliable. Put another way, if a test’s scores are not stable across time (i.e., if the test does not have test-retest reliability), then it cannot be valid as a measure of Intelligence. Even though validity often requires reliability, the reverse is not true. A measure might have excellent internal consistency and very high test-retest reliability, but we might not interpret it in a valid manner. In sum, a test must be reliable if it is to be interpreted validly, but just because a test is reliable does not mean that it will be interpreted validly.

The Importance of Validity

We hope that the importance of validity is apparent by this point in the chapter. We hope that the examples convinced you that validity is perhaps the most crucial issue in a test’s psychometric quality. In this section, we try to drive this point home by explicitly addressing the role and importance of validity in psychological research and practice. Whenever psychological measurements are conducted for any serious purpose, those measurements are meaningful only if
they have acceptable validity for their intended purpose. Without validity, those measurements are scientifically meaningless and potentially even hazardous.

Our ability to interpret the great bulk of behavioral research hinges on test validity. The goals of scientific research include describing, predicting, or explaining some aspect of our world – be it a physical or psychological aspect. Accurate description, prediction, and explanation depends on the ability to manipulate or measure specific variables that are deemed important. For example, some social psychologists have examined the hypothesis that exposure to violent video games increases one’s inclination to behave aggressively (e.g., Anderson & Dill, 2000; Bartholow, Sestir, & Davis, 2005). Research seems to indicate that, indeed, exposure to video violence does affect an individual’s aggressive behavior. But we must remain aware that this research partially hinges on the measurement of “inclination to behave aggressively.” If this key variable is measured with good validity, then we should have increased confidence in the conclusion that aggression is increased by exposure to video violence. However, if the “inclination to behave aggressively” is measured with poor validity, then we should have serious doubts about this conclusions. Without test validity, our understanding of the role of video games in teen aggressiveness is obscured.

Thus, validity influences the scientific process in a somewhat abstract sense, in that it affects the accuracy of our understanding of the world. Test validity can have an even more concrete impact on the outcome of the scientific process. As you know, another goal of scientific research is to guide decision-making about various aspects of our social world. Such decisions can be made at the societal level or at the individual level, and test validity has important implications for both kinds of decisions.
Without test validity, decisions about societal issues could be misinformed, wasteful, or even harmful. In June 2006, the U.S. House of Representatives held hearings concerning the level violence depicted in many popular video games. The underlying concern was that exposure to violent video games has dire consequences for those who play the video games, particularly for younger players. Specifically, the hearings were based on the assumption that exposure to violent video games increases aggressive behavior. Of course, the empirical legitimacy of this assumption is a key question. To the degree that there is strong research demonstrating the legitimacy of this assumption, then the Congressional hearings are on a relatively solid foundation. However, if the research is based on tests with poor validity, then we should have serious doubts about the meaning of the research. Consequently, any societal decisions based on such questionable research are themselves questionable. What would the implications be if the U.S. Congress enacted legislation on the basis of research with questionable validity? Congressional time and attention would be wasted, and misguided laws could be passed. Furthermore, it is conceivable that public funds could be spent to “prevent” aggression by intervening against violent video games. Again, if the research is flawed by poor test validity, then such funds would be wasted. Indeed, to the degree that public funds are diverted from alternative programs that actually would be beneficial to society, then wastefulness might even be harmful to people who would have benefited from those alternative programs. Let us clarify that we are not denigrating the quality of research in the effect of violent video games. Most of this research is indeed well conceived and well-executed. We simply use it to provide a compelling illustration of the fundamental connections between test validity, research quality, and social decision-making.
Finally, without test validity, test-based decisions about individuals could be misinformed or harmful. Decisions that are at least partially based on psychological testing include placing children in specific classes, selecting students for college admissions, hiring employees, making clinical decisions, and placing people in specific organizational roles. Such decisions have potentially life-altering implications for the individuals affected by them, and test validity can have an important impact on those decisions. To the degree that the decisions are based on well-validated measures, they hopefully benefit the test users and test takers. If decisions are based on the appropriate use of well-validated psychological tests, then (hopefully) children are more likely to be placed into appropriate classes, job applicants are more likely to be hired for jobs that fit their interests, skills, and abilities, students are more likely to be admitted to colleges that fit their academic skills, and clinical clients are diagnosed in ways that facilitate effective treatment. However, it is possible that such decisions are based on poorly-validated tests, or even on the inappropriate use of tests that have been well-validated for different uses. Recall the first chapter of this book, in which we discussed the North Carolina statute that “a mentally retarded person convicted of first degree murder shall not be sentenced to death” (session law 2001-346, senate bill 173). As we mentioned in that earlier chapter, the decisions regarding a person’s status as mentally retarded is based partially on the results from “an individually administered, scientifically recognized standardized intelligence quotient test administered by a licensed psychiatrist or psychologist.” We hope that the term “scientifically recognized standardized” is interpreted largely as a “scientifically validated” and we hope that lawyers are aware of this important issue.
Summary

This chapter presented the conceptual basis of test validity. As defined in the *Standards for Educational and Psychological Testing*, validity is “the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses” of a test (AERA, APA, & NCME, 1999, p. 9). We described several key implications of this way of thinking about validity – validity concerns the interpretation of test scores, it is a matter of degree, and it is based on empirical evidence and theory. Because empirical evidence is a key consideration in evaluating the validity of test scores, we also described five types of evidence that are relevant to test validity – test content, internal structure, response processes, associations with other variables, and the consequences of testing. We then contrasted the contemporary view of validity with traditional perspectives that are still commonly discussed and with reliability. Finally, we reiterated the importance of validity in terms of its implications for research and for real-world decision-making.
Table X-1

(Hypothetical) Correlation Matrix for a 2-Factor Set of Items

<table>
<thead>
<tr>
<th></th>
<th>Talkative</th>
<th>Assertive</th>
<th>Outgoing</th>
<th>Creative</th>
<th>Imaginative</th>
<th>Intellectual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talkative</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assertive</td>
<td>.66</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outgoing</td>
<td>.54</td>
<td>.59</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creative</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imaginative</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.46</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Intellectual</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.57</td>
<td>.72</td>
<td>1.00</td>
</tr>
</tbody>
</table>