PEARSON NEW INTERNATIONAL EDITION

Mastering Modern Psychological Testing
Theory & Methods

Cecil R. Reynolds Ronald B. Livingston

First Edition

ALWAYS LEARNING

PEARSON

Pearson New International Edition

Mastering Modern Psychological Testing
Theory & Methods
Cecil R. Reynolds Ronald B. Livingston
First Edition

PEARSON

- *Perceptual Reasoning Index*: Reflects perceptual and nonverbal reasoning, spatial processing, and visual-spatial-motor integration.
- *Working Memory Index*: Reflects working memory capacity and includes attention, concentration, and mental control.
- *Processing Speed Index*: Reflects ability to process nonverbal material quickly when adequate levels of attention are present as well as visual-motor coordination.

Factor Analysis: The Process. Factor analysis begins with a table of intercorrelations among the variables (individual items or subtests) that is referred to as a correlation matrix. Table 4 illustrates a correlation matrix showing the correlations between two subtests measuring memory (i.e., story memory and word list memory) and two subtests measuring visual-spatial abilities (visual analysis and visual-spatial design). A review of these correlations reveals that the two memory subtests showed moderate correlations with each other (i.e., $r = 0.52$), but small correlations with the visual-spatial subtests (ranging from 0.14 to 0.18). Likewise, the two visual-spatial subtests showed moderate correlations with each other (i.e., $r = 0.64$), but small correlations with the memory subtests (ranging from 0.14 to 0.18).

Several different factor analytic methods have been developed (often referred to as factor extraction techniques), and each has it own supporters and detractors. For example, principal component analysis (PCA) begins with the assumption that the variables have perfect reliability and places a value of 1.0 in the diagonal of the correlation matrix. With PCA all of the variance is analyzed, including shared variance (i.e., variance shared with other variables), unique variance (i.e., variance unique to individual variables), and error variance (i.e., variance due to measurement error). In contrast, principal factor analysis (PFA) does not assume the variables have perfect reliability and begins by placing the squared multiple correlation of all the other variables with the variable being considered (R^2) in the diagonal of the correlation matrix. With PFA, only shared variance is analyzed, with unique and error variance excluded. There are a number of other factor analytic methods available, but when the underlying factor structure is robust, it will typically emerge regardless of which extraction method is used—although the relative strength of the factors will vary across methods.

Several different factor analytic methods have been developed and each has it own supporters and detractors.

After selecting a factoring technique and applying it to the data, one must determine how many factors to retain. On one hand, the more factors retained the more variance that is explained

TABLE 4 Correlation Matrix

	Story Memory	Word List Memory	Visual Analysis	Visual-Spatial Design
Story memory	1.00	0.52	0.15	0.18
Word list memory	0.52	1.00	0.14	0.16
Visual analysis	0.15	0.14	1.00	0.64
Visual-spatial design	0.18	0.16	0.64	1.00

by the factor solution. On the other hand, retaining too many factors can result in a complex solution that does not facilitate interpretation. Selecting the number of factors to retain is not as straightforward a process as one might imagine, but researchers have a number of guidelines to help them. One common approach, referred to as the Kaiser-Guttman criteria, is to examine eigenvalues and retain values greater than 1.0. Eigenvalues reflect variance when each variable being analyzed contributes a variance value of 1.0. Retaining factors with eigenvalues greater than 1.0 ensures that each factor that is retained accounts for more variance than any single variable being analyzed. Another approach to determine how many factors to retain is the scree test (Cattell, 1966). Here factors are plotted on the horizontal axis and eigenvalues are plotted on the vertical axis. The researcher then examines the graph and looks for an "elbow," a point where previous factors explain substantially more variance than those past the point. Figure 4 presents a hypothetical scree plot. Examination of this plot suggests the presence of an elbow at the fifth factor. If you draw a line through the first five points you get a reasonably good fit. However, another line with a different slope would need to be drawn to fit the remaining points. Another important consideration is the interpretability of the factor solution. That is, does the factor solution make sense from a psychological perspective? Put another way, do the variables loading on a factor share a common theme or meaning? For example, on an intelligence test all the variables loading on a single factor might be measuring verbal processing skills. On a personality test, all the variables loading on a factor might reflect a propensity to experience negative affect. A factor solution that is not interpretable has little or no practical value and will likely provide scant evidence of validity.

All factor extraction methods produce a factor matrix that reflects the correlations between the variables and the factors (i.e., latent constructs). Table 5 presents a hypothetical factor matrix. In this example there are two subtests that measure memory-related abilities (i.e., story memory and word list memory) and two subtests that measure visual processing abilities

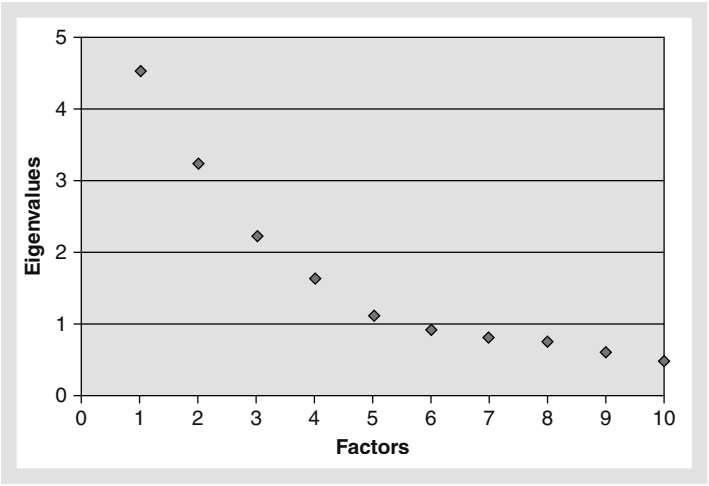

FIGURE 4 **Scree Plot.**

TABLE 5 Factor Matrix

Variables	Factor 1	Factor 2
Story memory	0.77	–0.49
Word list memory	0.74	–0.52
Visual analysis	0.62	0.61
Visual-spatial design	0.61	0.62

To enhance interpretability (i.e., understanding the meaning of the factors), most researchers geometrically rotate the axes of the factors.

(i.e., visual analysis and visual-spatial design). This initial factor matrix is difficult to interpret. All the subtests have moderate-to-high loadings on Factor 1. On Factor 2, the first two variables have negative loadings and the second two factors have loadings that approximate their loadings on Factor 1. To enhance interpretability (i.e., understanding the meaning of the factors), most researchers geometrically rotate the axes of the factors. The goals of these rotations are typically to eliminate any large negative loadings and to yield a solution where each variable has high loadings on only one factor and small loadings on all other factors (referred to as "simple structure"). If these goals are achieved, the rotated factor pattern should be easier to interpret. Table 6 presents a rotated factor pattern for the same subtests presented in Table 5. This rotated factor pattern is more interpretable, revealing that the two variables involving memory load on Factor 1, and the two variables involving visual processing load on Factor 2. Researchers have a number of options when selecting a rotation method. Some rotation methods produce orthogonal factors—the term orthogonal means the factors are not correlated. The most popular orthogonal rotation method is the Varimax technique. Other rotation techniques allow oblique factors—the term oblique as used here means the factors can be correlated. Most researchers select orthogonal rotations, because they are simpler to interpret, but oblique rotations may be appropriate when the factors are correlated in real-world applications (e.g., ability test batteries).

Confirmatory Factor Analysis. Our discussion of factor analysis to this point has focused on exploratory factor analysis. As described, exploratory factor analysis examines or "explores" a data set in order to detect the presence and structure of latent constructs existing among a set of variables. Confirmatory factor analysis is an alternative set of procedures that is gaining popularity among researchers. With confirmatory factor analysis the researcher specifies a hypothetical

TABLE 6 Rotated Factor Matrix

Variables	Factor 1	Factor 2
Story memory	0.90	0.06
Word list memory	0.90	0.11
Visual analysis	0.08	0.87
Visual-spatial design	0.09	0.87

factor structure and then examines the data to see if there is a reasonable fit between the actual and the hypothesized structure of the data set. There are a number of statistics available (referred to as model-fit statistics) that statistically test the fit or match between the actual and hypothesized factor structure. As Cronbach (1990) observed, a positive finding in confirmatory factor analysis does not necessarily indicate that the hypothesized structure is optimal, only that the data do not clearly contradict it. In summary, test publishers and researchers use factor analysis either to confirm or to refute the proposition that the internal structure of the tests is consistent with that of the construct being measured. Later in this chapter we will describe the results of both exploratory and confirmatory factor analytic studies of the Reynolds Intellectual Assessment Scales (RIAS; Reynolds & Kamphaus, 2003).

A positive finding in confirmatory factor analysis does not necessarily indicate that the hypothesized structure is optimal, only that the data do not clearly contradict it.

Factor analysis is not the only approach researchers use to examine the internal structure of a test. Any technique that allows researchersto examine the relationships between test components can be used in this context. For example, if the items on a test are assumed to reflect a continuum from very easy to very difficult, empirical evidence of a pattern of increasing difficulty can be used as validity evidence. If a test is thought to measure a one-dimensional construct, a measure of item homogeneity might be useful (AERA et al., 1999). The essential feature of this type of validity evidence is that researchers empirically examine the internal structure of the test and compare it to the structure of the construct of interest. This type of validity evidence traditionally has been incorporated under the category of construct validity and is most relevant with tests measuring theoretical constructs such as intelligence or personality.

Evidence Based on Response Processes

Validity evidence based on the response processes invoked by a test involves an analysis of the fit between the performance and actions the examinees actually engage in and the construct being assessed. Although this type of validity evidence has not received as much attention as the approaches previously discussed, it has considerable potential and in terms of the traditional nomenclature it would likely be classified under construct validity. For example, consider a test designed to measure mathematical reasoning ability. In this situation it would be important to investigate the examinees' response processes to verify that they are actually engaging in analysis and reasoning as opposed to applying rote mathematical algorithms (AERA et al., 1999). There are numerous ways of collecting this type of validity evidence, including interviewing examinees about their response processes and strategies, recording behavioral indicators such as response times and eye movements, or even analyzing the types of errors committed (AERA et al., 1999; Messick, 1989).

The *Standards* (AERA et al., 1999) noted that studies of response processes are not restricted to individuals taking the test, but may also examine the assessment professionals who administer or grade the tests. When testing personnel records or evaluating the performance of examinees, it is important to make sure that their processes or actions are in line with the construct being measured. For example, many tests provide specific criteria or rubrics that are intended to guide the scoring process. The Wechsler Individual Achievement Test—Second Edition (WIAT-II;

The Psychological Corporation, 2002) has a section to assess written expression that requires the examinee to write an essay. To facilitate grading, the authors include an analytic scoring rubric that has four evaluative categories: mechanics (e.g., spelling, punctuation), organization (e.g., structure, sequencing, use of introductory/concluding sentences, etc.), theme development (use of supporting statements, evidence), and vocabulary (e.g., specific and varied words, unusual expressions). In validating this assessment it would be helpful to evaluate the behaviors of individuals scoring the test to verify that the criteria are being carefully applied and that irrelevant factors are not influencing the scoring process.

Evidence Based on Consequences of Testing

Recently, researchers have started examining the consequences of test use, both intended and unintended, as an aspect of validity. In many situations the use of tests is based largely on the assumption that their use will result in some specific benefit (AERA et al., 1999; also see McFall & Treat, 1999). For example, if a test is used to identify qualified applicants for employment, it is assumed that the use of the test will result in better hiring decisions (e.g., lower training costs, lower turnover). If a test is used to help select students for admission to a college program, it is assumed that the use of the test will result in better admissions decisions (e.g., greater student success and higher retention). This line of validity evidence simply asks the question, "Are these benefits being achieved?" This type of validity evidence, often referred to as consequential validity evidence, is most applicable to tests designed for selection and promotion.

Researchers have started examining the consequences of test use, both intended and unintended, as an aspect of validity.

Some authors have advocated a broader conception of validity, one that incorporates social issues and values. For example, Messick (1989) in his influential chapter suggested that the conception of validity should be expanded so that it "formally brings consideration of value implications and social consequences into the validity framework" (p. 20). Other testing experts have criticized this position. For example, Popham (2000) suggested that incorporating social consequences into the definition of validity would detract from the clarity of the concept. Popham argued that validity is clearly defined as the "accuracy of score-based inferences" (p. 111) and that the inclusion of social and value issues unnecessarily complicates the concept. The *Standards* (AERA et al., 1999) appeared to avoid this broader conceptualization of validity. The *Standards* distinguished between consequential evidence that is directly tied to the concept of validity and evidence that is related to social policy. This is an important but potentially difficult distinction to make. Consider a situation in which research suggests that the use of a test results in different job selection rates for different groups. If the test measures only the skills and abilities related to job performance, evidence of differential selection rates does not detract from the validity of the test. This information might be useful in guiding social and policy decisions, but it is not technically an aspect of validity. If, however, the test measures factors unrelated to job performance, the evidence is relevant to validity. In this case, it may suggest a problem with the validity of the test such as the inclusion of construct-irrelevant factors.

Another component to this process is to consider the consequences of not using tests. Even though the consequences of testing may produce some adverse effects, these must be

contrasted with the positive and negative effects of alternatives to using psychological tests. If more subjective approaches to decision making are employed, for example, the likelihood of cultural, ethnic, and gender biases in the decision-making process will likely increase. This typically raises many controversies, and many professionals in the field attempt to avoid these issues, especially at the level of training students. We disagree. We believe this issue is of great importance.

INTEGRATING EVIDENCE OF VALIDITY

The *Standards* (AERA et al., 1999) stated that "Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use" (p. 9). The development of this **validity argument** typically involves the integration of numerous lines of evidence into a coherent commentary. As we have noted, different types of validity evidence are most applicable to different types of tests. Here is a brief review of some of the prominent applications of different types of validity evidence.

The development of a* validity argument *typically involves the integration of numerous lines of evidence into a coherent commentary.

- Evidence based on test content is most often reported with academic achievement tests and tests used in the selection of employees.
- Evidence based on relations to other variables can be either test-criterion validity evidence, which is most applicable when tests are used to predict performance on an external criterion, or convergent and discriminant evidence of validity, which can be useful with a wide variety of tests, including intelligence tests, achievement tests, personality tests, and so on.
- Evidence based on internal structure can be useful with a wide variety of tests, but has traditionally been applied with tests measuring theoretical constructs such as personality or intelligence.
- Evidence based on response processes can be useful with practically any test that requires examinees to engage in any cognitive or behavioral activity.
- Evidence based on consequences of testing is most applicable to tests designed for selection and promotion, but can be useful with a wide range of tests.

You might have noticed that most types of validity evidence have applications to a broad variety of tests, and this is the way it should be. The integration of multiple lines of research or types of evidence results in a more compelling validity argument. It is also important to remember that every interpretation or intended use of a test must be validated. As we noted earlier, if a test is used for different applications, each use or application must be validated. In these situations it is imperative that different types of validity evidence be provided. Table 7 provides a summary of the major applications of different types of validity evidence.

Although we have emphasized a number of distinct approaches to collecting evidence to support the validity of score interpretations, validity evidence is actually broader than the strategies described in this chapter. The *Standards* (AERA et al., 1999) stated: