CHAPTER NINETEEN

Psychometric Approaches to Understanding and Measuring Intelligence

SUSAN E. EMBRETSON AND KAREN M. SCHMIDT McCOLLAM

Since the *Handbook of Human Intelligence* appeared in 1982, the "psychometric approach" has changed dramatically. Traditionally, the psychometric approach was synonymous with the factor-analytic approach. Exploratory factor analysis was applied to discover the number and nature of the factors that underlie performance on cognitive tasks. Carroll's (1993) three-stratum model of intellect synthesizes the factors supported across hundreds of studies. Although the studies reported somewhat inconsistent factor patterns, Carroll found consistent support for several factors by reanalyzing their data with common methods of factor analysis.

However, the contemporary psychometric approach differs in three major ways from the traditional psychometric approach: (1) confirmatory approaches predominate over exploratory approaches, (2) structural analysis of items predominates over structural analysis of variables, and (3) item response theory (IRT) models predominate over factor analytic models. Thus, in the contemporary psychometric approach, confirmatory IRT models are applied to understand and measure individual differences. The intelligence construct is elaborated in confirmatory IRT models by comparing alternative models as explaining item responses. Some confirmatory IRT models include parameters to estimate the cognitive processing demands in items. These models permit items to be selected and banked by their cognitive demand features and provide results relevant to understanding what is measured by the items. Other confirmatory IRT models include parameters for person differences on the underlying processes, strategies, or knowledge structures. These models can define new types of individual differences. As will be elaborated below, parameters are included to measure qualitative differences in item responses such as relative success in various underlying cognitive operations, use of different strategies or knowledge structures, and modifiability of ability with intervention.

In this chapter, we will first describe the major historical exploratory factor analytic theories and their implications for measuring and understanding intelligence. Then, several IRT models will be elaborated. Both unidimensional and multidimensional models will be presented. For each IRT model, we will present an overview of the model, describe some key applications, and present one or two elaborated examples to illustrate the potential of the model.

FACTOR ANALYTIC APPROACHES TO MEASURING AND UNDERSTANDING INTELLIGENCE

Factor analysis has been the primary tool for measuring and understanding intelligence for several decades. In this section, we begin with the historical foundations of measurement through factor analysis. Factor analysis has not only been important for understanding the intelligence tests that have been developed but has also provided a rationale for several testing methods. Factor analysis remains influential in testing.

Exploratory Factor Analysis Methods

General Factor Emphasis

SPEARMAN. Spearman (1904) proposed a two-factor theory of intelligence in which performance was determined by a general factor (g), a universal due to a person's general intelligence, and a specific factor (s) due to a unique ability or activity related to a particular test. Spearman (1904) suggested "all branches of intellectual activity have in common one fundamental function (or group of functions), whereas the remaining or specific elements seem in every case to be wholly different from that in all the others" (p. 202). Although both factors are present within each intellectual activity, their relative weight varies from activity to activity (Spearman, 1904, 1927). For example, Spearman's "abilities," originally defined by school subjects, had a greater relative g-to-s ratio for classics than for music study (Spearman, 1927).

On the basis of preparatory school student performance data, Spearman (1904) found an overall pattern of correlations among all of the various intellectual activities, indicating uniformity. The general factor is responsible for two tests being correlated. However, Spearman's two-factor theory holds only if a test battery includes only one of each type of test. Spearman (1927) prescribed this universality of g for application to the measurement of individual differences. For example, by measuring an individual's abilities on a series of tests, g can be determined, explaining much information about some of the abilities, and some about all abilities. Supplemental performance variation would be explained by s. In addition, Spearman's (1927) suggestion for test design was governed by the extent to which the test measures an individual's g or s.

TETRAD DIFFERENCES. Spearman's method of tetrad differences was the result of his observation that the true difference between correlation products of different abilities equals zero. The form of the equation is as follows:

$$(r_{ab} \times r_{cd}) - (r_{ac} \times r_{bd}) = 0.$$

For example, suppose the following four tests, a = French; b = English; c = Music; d = Math, and their correlations are as follows: $r_{ab} = .750$; $r_{cd} = .500$; $r_{ac} = .600$; $r_{bd} = .625$. Of course, the true difference and the observed difference are not always the same owing to sampling error. If tests are correlated through the specific factors, for example for two memorizing symbols tests, then the tetrad equation becomes invalid.

The meaning of g for Spearman was "objective" because of this tetrad equation. In fact, Spearman (1927) stated, "Eventually, we may or may not find reason to conclude that g measures something that can appropriately be called 'intelligence.' Such a conclusion, however, would still never be the definition of g, but only a 'statement about' it" (p. 76). He originally proposed the psychological meaning of g to be mental energy, concentration, or will power. For the physiological meaning, he hypothesized that neural plasticity or neural energy was important for g (Spearman, 1927). Later, however, Spearman included "agreement and difference" in his psychological characterization of g. For example, the presence of a spatial factor and a g factor together also implies that the absence of a spatial factor entails the absence of g (Spearman & Jones, 1950).

EVIDENCE FOR g. Spearman (1927) presented a variety of evidence concerning g in psychological tests that embody different relationships. Summarizing the correlational patterns of tests with tetrad differences, Spearman concluded that g exists in all types of relationships to the same degree and that cases in which "group factors" contribute to the correlations between tests are rare. However, a closer examination of Spearman's data reveals that the patterns of correlations did not support the two-factor theory so strongly as Spearman's conclusions would indicate. Although Spearman elaborated 10 types of relationships in tests, which he categorized as ideal types (evidence, likeness, and conjunction) and real types (space–time, attribution, identity, constitution, causality, and psychological), evidence was obtained for only 5 relationships. Four of the five types were moderately saturated with g comparable to factor loading magnitudes in the .70s. The remaining type of relationship (time) only weakly supported g with a saturation comparable to a factor loading in the .30s. Further, for only one type of relationship did the evidence clearly support only one common factor. Weak evidence for group factors was obtained for two types of relationships, whereas strong evidence was obtained for two

other types of relationships. The inconsistencies in Spearman's own data on the two-factor theory foreshadowed what was to come, namely, the identification of significant group factors.

SPEARMAN'S COGNITIVE THEORY. The general factor *g* in Spearman's psychometric theory is not really an explanatory concept. It is simply a description given to the central factor in a battery of tests or item types. Spearman (1923) also proposed a cognitive theory to explain intelligent thought. The analogy item played a central role in Spearman's cognitive theory, and defined *g* (i.e., it was highly saturated with *g*). Spearman postulated three qualitative principles of cognition to account for intelligent behavior, which he called "noegenetic" thinking. Spearman regarded the analogy as the paradigm of noegenetic thinking because solving analogies depends on all three principles.

Consider first the three principles as elaborated by Spearman (1923). The first principle is *apprehension of experience*. As explained by Spearman (1923, p. 48), "Any lived experience tends to evoke immediately a knowing of its direct attributes and its experiences." Stimuli are meaningful for persons when they have relevant experiences or knowledge of related attributes.

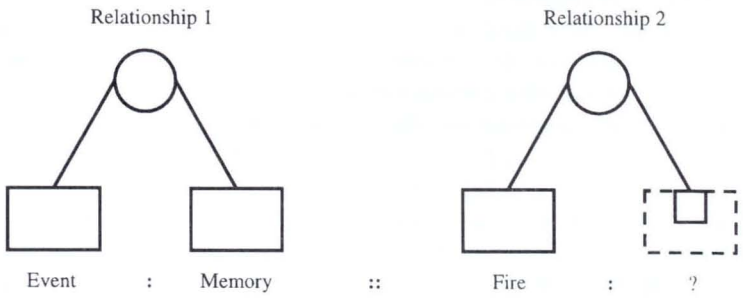
The second principle is the *eduction of relations*. According to Spearman (1923, p. 63), "The presenting of two or more characters tends to evoke immediately a knowing of relation between them." The second principle involves inference. The third principle is the *eduction of correlates*. "The presenting of any character together with a relation tends to evoke immediately a knowing of the correlative character" (Spearman, 1923, p. 91). That is, given a new stimulus and a relationship, a new stimulus that fulfills the relationship can be anticipated.

Sternberg's (1977) theory expanded Spearman's three principles and connected them to contemporary information-processing theory. The first principle is simply an assurance that organisms are able to understand their environment and act intelligently within it. The contemporary cognitive psychology process of encoding is similar to the apprehension of experience. The second and third principles concern the way in which objects, characters, or fundament in the environment are cognitively organized and processed. The second principle, the eduction of relations, is similar to Sternberg's (1977) concept of inference in which a relationship is inferred between pairs of stimuli. The third principle, the eduction of correlates, is similar to Sternberg's (1977) application in which a new stimulus is anticipated to fulfill a relationship that is applied to a stimulus.

To understand how the two processes of eduction are presented in prototypic form in the analogy item, it is necessary to compare the theory with the structure of the analogy item. Figure 19.1 presents a general schematization of Spearman's second and third principles of cognition along with an analogy item. In the part of the drawing labeled Relationship 1, the boxes represent two fundaments, which are given. The left side of the analogy contains two fundaments from which some relationship can be educed. The educed relationship is represented by the circle above the fundament. The right side of the drawing, Relationship 2, represents the eduction of correlates in which one fundament and the relationship are given. Now, for the analogy item below, it can be seen that the only part given on the right side is the fundament. Solving the item then depends on using the relationship educed on the left to educe the correlate on the right side.

If Spearman's (1923) theory is to account for a broad range of behavior, it should be apparent that the fundaments should incorporate a broad range of phenomena. Spearman indicated that the fundaments are not restricted to simple elements. The relationship between two fundaments may define a higher order fundament which, in turn, can be

FIGURE 19.1. Spearman's cognitive theory applied to verbal analogies.



related to another higher order fundament. The relationship between higher order fundaments, so derived, would be a relationship of relationships. Neither the number of levels from which the fundaments are removed from actual objects nor the range of objects has any limits, according to Spearman.

Spearman's cognitive theory gives the *g* concept explanatory power. Of course, the status of Spearman's cognitive theory is questionable because it is quite old and relatively untestable with the methods available to Spearman. Sternberg's (1977) contemporary theory of information processing on analogies incorporated many principles similar to Spearman's cognitive theory. Sternberg's (1977) theory was supported by modern methods of mathematical modeling (see also Sternberg, 1985).

Multiple Factor Emphasis

Other researchers disagreed with Spearman's theory and method. For example, Holzinger and his colleagues (Holzinger & Swineford, 1939; Holzinger & Harman, 1941) found conditions in which the two-factor theory was not applicable, namely, bifactor situations, which they described as both general and secondary group factors necessary for reproducing correlations in complex sets of variables. Others, such as Hotelling (1933) and Kelley (1935) emphasized different methods for extracting factors, such as principal components analysis, in which the first axis has the best possible fit to the entire matrix of scores, and the second axis has the next possible fit, lying perpendicular to the first axis.

KELLY. Kelly (1928), who was interested in the method and content of group factors, carefully analyzed test performance in several age groups. Kelly thoroughly examined test pair bonds to determine their shared nature and to build factors uncontaminated by other factors in order to obtain independent mental trait factors. In seventh-grade, third-grade, and kindergarten populations, Kelly identified verbal facility, number facility, memory, spatial facility, and interest factors, and he identified general and speed factors for the seventh-grade population. Kelly (1928) defined intelligent behavior as "largely the regulation of impulses and the co-operation between adaptability and persistence, while intellect may involve a more abstract analytical capacity besides" (p. 223).

THURSTONE. Thurstone (1938) proposed a theory of primary mental abilities in which 9 independent factors were identified by examining 250 select college students on 56 tests: memory, number, verbal comprehension, induction, deduction, arithmetic reasoning, word fluency, space, and perceptual speed. Later, Thurstone (1941) examined high school students and found the same factors present. In another sample of 700 eighth-grade children, Thurstone (1941) found most of the original factors present, although the factors were correlated in this young age group. In addition, Thurstone believed the obtained second-order general factor was "probably the general factor which Spearman has so long defended" (Thurstone, 1941, p. 111). In addition, verbal comprehension, reasoning, and induction factors had the highest second-order general factor loadings.

GUILFORD. Guilford's structure of intellect model (Guilford, 1967, 1977) is a three-facet model that explains mental operations, stimulus content, and response forms or products. The model contains up to 150 different abilities. Guilford (1967) rejected a general factor. Thirteen studies from an aptitudes project on young adults showed about 17% of all test variable correlations being at or near zero.

Guilford (1948) valued factor analysis in developing tests for personnel selection and classification because it can result in minimally correlated scores that represent different contributors for predicting success. Additionally, Guilford (1948) believed that factor analysis should be applied in a planned experimental investigation rather than as an afterthought for a set of convenient intercorrelations. Guilford (1985) typically extracted a larger number of factors than indicated by the SCREE criterion to make better use of information and to avoid ignoring a potential factor.

HUMPHREYS. Humphreys' (1985) theory of general intelligence was broadly defined. His tests included heterogeneous item types from different varying dimensions. He believed that heterogeneous items preserved predictive validity over reliability. Especially appealing to Humphreys for measuring intelligence were items that shared a particular attribute, such as content, but loaded on different factors. Humphreys' goal was to include one dominant

dimension rather than one purely defined, homogeneous dimension. Humphreys' theory of general intelligence measurement included several measurement principles, as follows:

1. Items can be added together to form a total score if they are positively correlated.
2. Items should be as heterogeneous as possible within certain defined parameters.
3. A test should be broadly defined without losing its measurement purpose.
4. Psychometric analyses should be applied to check assumptions but not to make decisions in test construction.
5. The item homogeneity criterion should be used sparingly; tests can be rejected for either too much or too little homogeneity.

Humphreys (1952) believed that abilities should be organized in a hierarchical structure. For example, a set of tests measuring mechanical information about tools would be organized into three different levels: (1) narrow tool group factors, (2) broad mechanical area factors, and (3) a general mechanical information factor. Using orthogonal factor transformation techniques to create a more parsimonious matrix allows one to interpret factors at all levels. However, Humphreys suggested that interpretation be limited to the broad and general levels.

VERNON AND BURT. Vernon (1950) defined ability as *g* at one level of a hierarchy and then two broad abilities at the next level: *v:ed* (verbal–educational) and *k:m* (practical–mechanical). Under *v:ed*, verbal and numerical specific abilities are factored, and under *k:m*, spatial and mechanical abilities are factored. Vernon (1961) cautioned that these group factors are infinitely divisible, depending upon the level of detail at which the analysis is performed. A similar hierarchical structure was also proposed by Burt (1949).

CATTELL AND HORN. Cattell and Horn (Cattell, 1963; Horn & Cattell, 1966) organized abilities according to a hierarchical structure as well with *g* divided into major factors, fluid intelligence (*gf*) and crystallized intelligence (*gc*). Fluid intelligence was hypothesized to be more genetically influenced and based on physiological aspects of an individual. Crystallized intelligence was hypothesized to be experientially determined or acculturated. Tests of fluid intelligence include response to novelty, problem solving, and figural reasoning; tests of crystallized intelligence include, most prominently, knowledge-based tests such as vocabulary and mathematical tests.

CARROLL. Carroll's three-stratum theory of intelligence (Carroll, 1993) was developed through surveying and factor analyzing over 460 prominent datasets in the literature. The stratum theory is also hierarchical. Stratum 1 comprises narrow factors, which are first-order factors to explain the correlation matrix of test items. Stratum 2 comprises broad factors that explain the correlations of the Stratum 1 factors. Stratum 3 is a general factor *g*, which explains the correlations of the Stratum 2 factors. Carroll's theory extends the theories of Thurstone (1938), Guilford (1967), and Horn and Cattell (1967). However, the three-stratum model also includes several narrowly defined abilities, such as phonetic coding and perceptual illusions, that typically are not included in such models.

The 1921 Symposium on Intelligence and Its Measurement

After Spearman introduced his two-factor theory and others followed with alternatives, a movement for discussion of conceptions of intelligence occurred. Seventeen leading researchers gathered in 1921 to discuss the definition of intelligence and its measurement. Most investigators disavowed the notion of general intelligence and desired that its definition be expanded. It should be noted that Spearman, who was the primary advocate of general intelligence, did not attend the conference.

Among the most outstanding comments made by the investigators were R. L. Thorndike's statement: "The value of a test score is its value in prophesying how well a person will do in other intellectual tasks" (1921 symposium, p. 125). Specifically, Thorndike called for using zero-order and partial correlations of simple and analytical processes with the criterion task to understand their explanatory contributions better.

Terman's view of intelligence involved grasping the significance of adaptive situations and that "An individual is intelligent in proportion as he is able to carry on abstract thinking" (1921 symposium,

p. 128). Terman critiqued the notion of a singular type of test measuring a mental function with the following comments: "...a test does not, at all points, bring the same kind of mental activities into play. Success in the easier part of the test may depend chiefly on the subject's ability to remember what he is told to do; see likeness and differences on the representative level, or even to a considerable extent on eye–hand coordination in the use of a pencil" (1921 symposium, p. 131). Hence, Terman viewed intelligence as multifaceted and operating jointly at different levels.

Colvin suggested that "An individual possesses intelligence in so far as he has learned, or can learn to adjust himself to his environment" (1921 symposium, p. 136). He recommended that testing proceed by focusing on analysis, synthesis, and attention span and disregarding speed. In contrast, Thurstone's definition of intelligence contained three components: (a) inhibitive capacity, (b) analytical capacity, and (c) perseverance. Inhibitive capacity involves substituting instinctive, environmental, and social pressure with conceptual thinking. Analytical capacity involves inhibiting instinctive pressure to ensure response flexibility. Perseverance involves volitional energy.

Perhaps most striking is the diversity of definitions and qualities of intelligence noted by these researchers. A much more recent survey of researchers in intelligence (Sternberg & Detterman, 1986) found similar divergence of views and many similar beliefs. Overall, the researchers agreed upon continued investigations for the improvement of measurement and expansion of knowledge about intelligence. It is noteworthy that decades of research have not led to convergence of theoretical views, however.

Confirmatory Factor Analysis

Joreskog and Sorbom's confirmatory factor analytic work has made an enormous impact on the field of psychometrics (Joreskog, 1969; Joreskog & Sorbom, 1989). Their LISREL (Joreskog & Sorbom, 1989) methods have allowed researchers to investigate hypothesized models in relation to sample data and alternatives that explain the factor structure of test intercorrelations and underlying processes. Confirmatory factor analysis computer programs are growing in number and becoming more widely available, including EQS (Bentler, 1985), AMOS (Arbuckle, 1993), and Mx (Neale, 1994), allowing for widespread use among social science investigators, including those interested in measuring intelligence constructs.

Confirmatory factor analysis is now applied widely to study intelligence. In fact, it is usually preferred over exploratory factor analysis owing to the massive literature on theoretical factor structure. To illustrate the advantages of confirmatory factor analysis, consider Kyllonen and Christal's (1990) study of the relationship of general intelligence (reasoning) to working memory capacity. In an effort to understand processing, content, and methodological aspects of reasoning and working memory for delineating their relationship, Kyllonen and Christal (1990) examined multiple tests describing these constructs in four large-sample studies. The authors carefully considered experimental and correlational disciplines for defining factors and included speed and knowledge factors in addition to working memory and reasoning. In addition, within reasoning and working memory, they analyzed domain-specific and domain-independent aspects of tests for describing the nature of the factors.

For Study 1, working memory and reasoning factor correlations ranged from .79 to .93 across fitted models. Working memory was found to be general in nature, for both linguistic and quantitative processes of working memory were correlated. For Study 2, testing modalities and domains were varied. The reasoning factor's tests were broadened to eliminate confounding with working memory content, and computer-administered tests were given within both factors. In addition, two reasoning factors were defined to understand method variance better. Fit analyses indicated better fit with the two-factor reasoning model over the one-factor reasoning model. Reasoning was found to be more highly related to knowledge, and working memory with speed. For Study 3, a more broadly defined reasoning factor was defined by including ETS Kit tests (Ekstrom et al., 1976), and similar results were found. For Study 4, the reasoning factor was defined more generally to detect any changes in its correlation with working memory. Across all four studies, the correlations between working memory and reasoning were .82, .88, .80, and .82, respectively, suggesting that differences in working memory capacity are strongly

related to executive processes regardless of variation in test content and administration method.

The obvious strengths in using confirmatory factor analysis in Kyllonen and Christal's (1990) approach are (1) multiple tests defining the constructs can be fitted and compared using hypothesized and alternative models, (2) a variety of test content within constructs can be used to test domain specificity, and (3) method variance can be better understood by using different types of administration techniques. The major weakness is that constructs are clearly defined between rather than within factors.

Implications for Measuring and Understanding Intelligence

The factor analytic approaches have had major impact on how intelligence has been measured and understood. We will first elaborate impact on measurement and then elaborate impact on understanding.

Measuring Intelligence

Neither Spearman's (1927) psychometric theory nor Spearman's (1923) cognitive theory was uniquely associated with a test of intelligence. A single-factor test, which includes only a single item type that is highly saturated with *g*, is most consistent with both Spearman's psychometric theory and cognitive theory. Given the central role of analogies in both theories, one would have expected an analogy test to be developed directly from Spearman's theories. Surprisingly, however, the single factor test that is most closely associated with Spearman consists of matrix problems. John Raven, Spearman's student, developed a matrix completion test that reflects Spearman's principles of eduction of relation and eduction of correlates. Raven's matrix problems (see Raven, 1956) consisted of a three-by-three array of complex figures that varied systematically across the rows and columns. The last element was missing. To find the missing element, relationships among elements had to be educed, and then a correlate to complete the missing entry had to be educed. Spearman is reported to have favored matrix problems for educing relationships, and he had huge displays of matrix problems in his office. Tests with Raven's matrices have been used cross-culturally for several decades to measure general reasoning. The Raven's Progressive Matrices Tests have recently been updated and remain an important measure of intelligence (Raven, J., Raven, J. C., & Court, 1995).

However, a general intelligence test, with mixed item types, is compatible with Spearman's two-factor theory as well. Because *g* was postulated to be the only source of correlations between distinct types of items, the dimension measured from a mixed test would be *g*. The influence of specific factors, balanced over item types, would essentially be cancelled. According to Spearman's view, a test of heterogeneous items, with varying saturations with *g*, would be inefficient but would still measure *g*. However, Humphreys' views about item heterogeneity would be well represented by a test of mixed-item types. In any case, many intelligence tests employ mixed-item types. Even the first successful intelligence test, Binet's scales of intelligence, is compatible with these views. The current revision of the Stanford–Binet still renders a global score from mixed-item types (Thorndike, Hagen, & Sattler, 1986). Similarly, many contemporary intelligence tests, such as the Wechsler scales (WAIS–R; Wechsler, 1981; WISC–III; Wechsler, 1991) continue to provide an overall index of intelligence based on mixed-item types.

In contrast, several tests have been directly associated with the multiple factor theories. For example, the Primary Mental Abilities Test (Thurstone & Thurstone, 1941) resulted directly from Thurstone's application of multiple factor analysis. More recently, the Schaie–Thurstone Adult Mental Abilities Test (Schaie, 1985) has updated the original Primary Mental Abilities Test. Similarly, Guilford (1967) developed a battery of tests to represent factors in his theory. In other cases, tests that were developed later were inspired by factor theories. Examples of multiple aptitude tests of separate abilities include the ETS Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, Harman, & Dermen, 1976), the Differential Aptitudes Test (Bennett, Seashore, & Wesman, 1984), and the Armed Services Vocational Aptitude Battery (Moreno, Wetzel, McBride, & Weiss, 1984).

Many contemporary intelligence tests have been influenced by the hierarchical theories of intellect. Scores at different levels are routinely reported in contemporary tests. For example, the Kaufman Adolescent and Adult Intelligence Test (Kaufman & Kaufman, 1993), the Woodcock–Johnson Tests of

Cognitive Ability (Woodcock & Johnson, 1989) and the Differential Ability Scales (Elliott, 1990) provide scores for fluid and crystallized intelligence.

Both exploratory and confirmatory factor analysis are important in determining the scores reported from intelligence tests with mixed-item types. For example, on the Differential Ability Scales (Elliott, 1990), the actual structure of the abilities to be measured at various ages was determined by confirmatory factor analysis.

Understanding Intelligence

Factor theories have also had an impact on how intelligence is understood. Factors are derived from the similarity of persons' responses to items or subtests. Items that are responded to in the same way load on the same factor or factors. The factor analytic method identifies factors that summarize patterns of correlations. The factors are often interpreted as latent variables that underlie test or item performance. The nature of the dimension typically is determined by an inspection of the items that load on it. That is, the dimension is interpreted by comparing features shared by the items.

For a single-factor theory, such as postulated by Spearman (1927), item features are quite diverse because all cognitive tasks depend on *g* to some extent. Thus, the nature of *g* must be determined from other considerations. The most prevalent approach has been to correlate measures of *g* with external variables such as group membership, other tests, criterion variables, and so forth to build a nomological network according to Cronbach and Meehl's concept of construct validity. Thus, the theoretical meaning of *g* depends on its empirical relationships. Another approach is to develop a theory about the item types that load highly on *g*. For example, Spearman (1923) postulated processing mechanisms to explain performance on verbal analogies, an item type that was highly saturated with *g*. The latter approach, termed the construct representation aspect of construct validity (Embretson, 1983), has been applied effectively only in the last two decades (e.g., Sternberg, 1977). Further, the processing mechanisms underlying the items that appear on existing tests have rarely been examined empirically.

For multiple factor theories, items or subtests that load on the same factor often share obvious features such as verbal comprehension or spatial visualization. These shared features give rise to interpretations. However, the shared features are often quite global so that further understanding must be provided from outside considerations like understanding *g*. The most prevalent approach has been to elaborate empirical relationships with other variables. Confirmatory factor analysis (e.g., Kyllonen & Christal, 1990), along with structural equation modeling, has been particularly effective in expanding nomological networks for multiple factors. However, the mechanisms underlying item performance are not necessarily elaborated by empirical correlates with other variables. Unless the other variables represent theoretically singular dimensions, it is difficult to pinpoint the source of the correlation.

ITEM RESPONSE THEORY MODELS FOR MEASURING AND UNDERSTANDING INTELLIGENCE

Both exploratory and confirmatory IRT models are available to measure and understand intelligence. The exploratory models were developed earlier and have been applied more extensively. Unidimensional and multidimensional exploratory IRT models have been applied. We will begin by presenting exploratory models. Although the exploratory IRT models have substantial practical advantages, applying them usually results in measuring aspects of intelligence that are similar to the aspects derived from the factor analytic approaches. Yet, it is valuable to elaborate these models for two reasons: (1) IRT models and their many properties, are unfamiliar to many psychologists and (2) the advantages of the confirmatory models depend directly on the properties of IRT models.

Unidimensional Item Response Theory Models

Item response theory is rapidly replacing classical test theory as the psychometric basis for testing and has many theoretical and practical advantages over classical test theory. The exploratory models mentioned below have become quite routine in testing. Some typical models and some elaborated applications will be discussed to illustrate some properties. More extended treatments are available in textbooks (see Embretson & Reise, in press;

Hambleton, Swaminathan, & Rogers, 1989). The exploratory models do not permit direct incorporation of theory into the estimates that are obtained. The confirmatory IRT models, however, were designed to incorporate theory explicitly in the measurement process.

Exploratory IRT Models

Some Logistic IRT Models

In IRT, persons are measured in the context of a model of the item response process. IRT models are mathematical models of the responses of persons to particular tasks or items. Thus, ability estimates depend not only on persons' responses but also on the properties of the items.

IRT models contain parameters to represent the characteristics of items and persons. The probability that a person solves a particular item depends jointly on ability level and on the item characteristics. In typical IRT models, ability level and item difficulty combine additively to produce the probability that the item is endorsed or passed. Many IRT models are based on the logistic distribution in which the parameters are exponents. For example, the Rasch (1960) model predicts item success from the simple difference between the item's difficulty b_i and the person's ability θ_j as follows:

$$P(X_{ij} = 1 \mid \theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \qquad (19.1)$$

Abilities and item difficulties that are estimated from Equation 19.1 are similar in magnitude to z scores. High values are assigned to high abilities and difficult items.

More complex unidimensional IRT models add parameters to reflect additional item properties. Items, for example, may differ in how well they discriminate between levels of ability and in their vulnerability to guessing. Thus, the simple model in Equation 19.1 can be expanded to include item discrimination a_i and guessing c_i as follows:

$$P(X_{ij} = 1 \mid \theta_j, b_i, a_i) = c_i + (1 - c_i)\frac{\exp(a_i(\theta_j - b_i))}{1 + \exp(a_i(\theta_j - b_i))} \qquad (19.2)$$

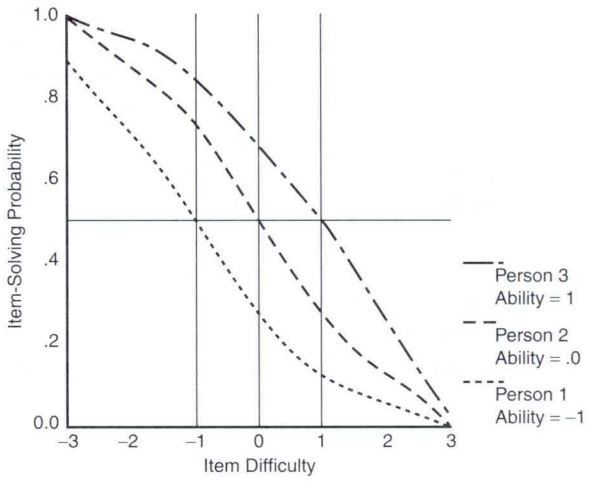
Notice that the impact of the difference between the person's ability and the item's difficulty is proportional to item discrimination, which is the multiplier a_i in Equation 19.2. Notice also that the guessing parameter c_i prevents the response probability from falling to zero, even for the lowest ability levels.

Meaning of the Parameters

Three types of meanings for the parameters estimated from Equation 19.1 or Equation 19.2 may be illustrated: (1) the impact of ability on performance, (2) the impact of item difficulty, and (3) conjoint scaling of persons and items. These meanings are essential to the advantages of IRT over classical test theory. In the examples below, a 30-item test of fluid ability, the Abstract Reasoning Test (ART; Embretson, 1995), was calibrated with the Rasch model (i.e., Equation 19.1) for a sample of 818 young adults.

MEANING OF ABILITY FOR PERFORMANCE. The meaning of a person's ability for item responses is easily shown by *person characteristics curves* (PCCs). Suppose that three persons are selected from the sample with abilities of −1.0, .00, and 1.0 (similar in magnitude to z scores), respectively. By inserting item difficulties at various levels in Equation 19.1, a probability for solving each item may be calculated. Figure 19.2 shows the PCCs for the three persons. Item difficulty is represented on the abscissa. Item solving probability is represented on the ordinate. The point of inflection occurs at the probability of .50, which is shown by a reference line from the ordinate axis on Figure 19.2. The item difficulty for the point of inflection is analogous to

FIGURE 19.2. Person characteristics curves.



a threshold value. Like a psychophysical threshold, items at a threshold are as likely to be passed as to be failed.

Several meanings for an ability level are shown in Figure 19.2. First, items at the person's threshold may be identified. If the difference between ability and item difficulty is zero, the probability of passing is .50. Three reference lines through the abscissa at the item values that correspond to the three abilities, −1.00, .00, and 1.00, are shown in Figure 19.2. Notice that the reference lines intersect the line through a probability of .50 differently for each PCC. For the PCC at the ability of −1.00, for example, the reference line through the ordinate probability of .50 crosses the reference line from the item difficulty at −1.00. Thus, one meaning for ability is the scale value of items at the person's threshold. Second, diagnostic information about the person's relative success on other items may readily be obtained. If a person's ability exceeds item difficulty, then the difference is positive and Equation 19.1 predicts a probability of success greater than .50. Conversely, if the person's ability falls below item difficulty, then the probability of passing is less than .50. For Person 2, for example, items with difficulties less than .00 are relatively easy, whereas items with difficulties greater than .00 are harder.

Of course, the actual response of the persons to the 30 ART items is known. In what ways do PCCs provide additional information? If the IRT model fits the data, the advantages of the PCC for diagnosing performance include: (1) more detailed descriptions of performance than actual item responses because the latter has only two values, pass or fail; (2) more accurate description of performance because error factors may result in the person's actually failing an easy item and passing a hard item; (3) predictions for items that the person has not been presented, if their item difficulties are known; and (4) possible detection of aberrant response patterns because fit of the person to the model can be checked by comparing actual responses to predictions.

ITEM CHARACTERISTICS CURVES. Diagnostic information about items also can be given by Equation 19.1 predictions. An *item characteristics curve* (ICC) regresses item solving probabilities on ability level. Figure 19.3 shows ICCs for three ART items from Table 19.1. Like the PCCs, the ICCs

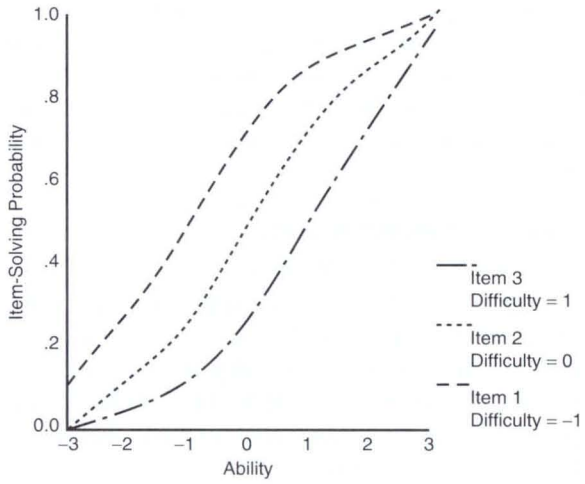


FIGURE 19.3. Item characteristics curves.

are S-shaped. In the middle of the curve, large changes in item-solving probability are observed for small ability changes, whereas at the extremes, item-solving probabilities change very slowly with ability changes.

The three ICCs in Figure 19.3 have the same shape but differ in location. Location is indicated by item difficulty; it is the inflection point at which the item-solving probability is .50. Like the PCCs, location is directly linked to ability level. For example, the ability level that has a probability of .50 for solving Item 1 is −1.00, which is its item difficulty.

In Figure 19.3, all items have the same discrimination and a lower asymptote of zero. If the three-parameter logistic model in Equation 19.2 had been applied to the ART data, items would have differing slopes (item discrimination) and a nonzero lower asymptote (due to guessing).

CONJOINT SCALING. Conjoint scaling means that item difficulty and person ability are placed on a common scale. Figure 19.4 presents a joint frequency distribution of item difficulty and ability on the ART. Notice how items are located on the same scale as persons. Further, the distributions may be compared to determine if the test is appropriately targeted to the sample. In Figure 19.4, item difficulties have the highest frequency near the middle of the ability distribution; thus, most items are appropriate for most persons. However, Figure 19.4

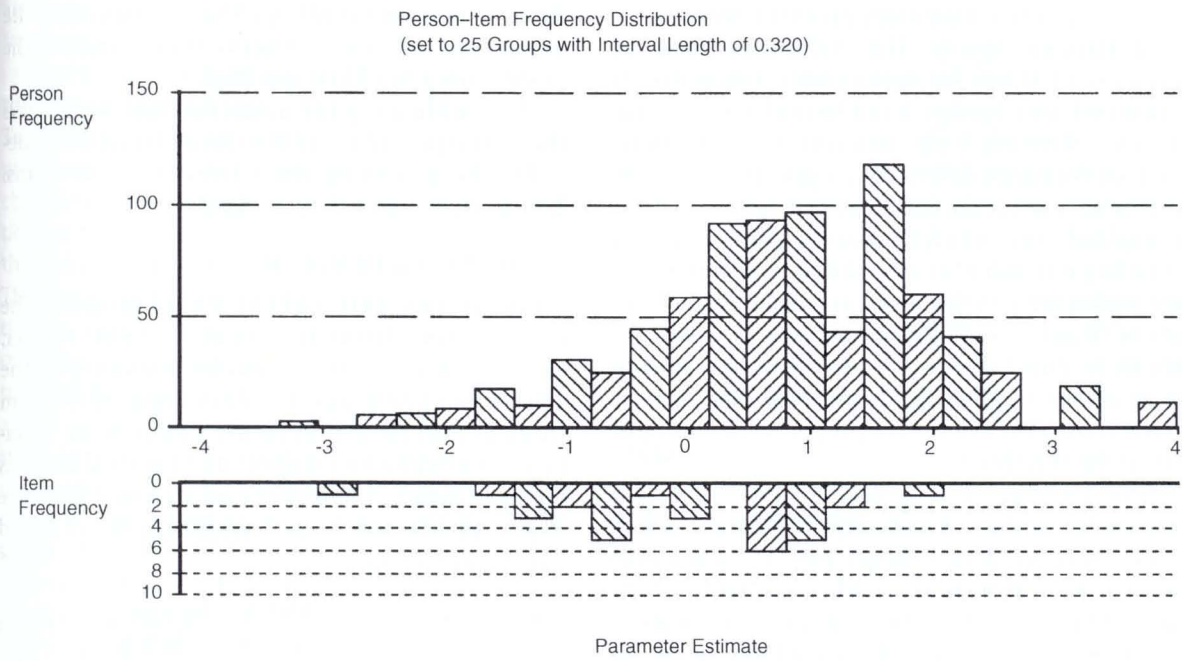


FIGURE 19.4. Conjoint distributions of persons and items.

also shows a noticeable lack of difficult items and many persons with high abilities. Thus, more difficult items are needed to measure these persons optimally.

ADVANTAGES. The theoretical advantages of IRT models include the following:

1. Justifiable interval or ratio-level scaling of persons.
2. Population-invariant item calibrations.
3. Item-invariant meaning for ability.
4. Item-referenced meaning for ability levels.
5. Superior capability for handling missing data (e.g., persons do not receive all the same items).
6. Standard errors of measurement that reflect the appropriateness of items for the various ability levels.

Each of these advantages will be elaborated in turn.

First, conjoint scaling means that persons and items are located on the same continuum, as shown in Figure 19.4. The same increase in item probabilities results from either increasing ability or decreasing item difficulty. Andrich (1988) pointed out that the conjoint additivity criterion for fundamental measurement is met directly by successful scaling with the Rasch model. Thus, interval-level scaling can be justified by fitting the Rasch. In contrast, classical test theory can be justified as interval-level scaling under only limited conditions (see Embretson, 1996). A practical advantage that derives from interval-level scaling is measuring individual change to reflect treatment or developmental effects. When interval level scaling is not obtained, as is true for tests developed under classical test theory, change scores have paradoxical reliabilities, spurious correlations with initial scores, and nonmonotonic relationships to true change (see Bereiter, 1963). Failing to achieve interval scaling also influences means and variances, which will lead to biased estimates of effects and their significance (see Embretson, 1995; Maxwell & Delaney, 1985). Second, population-invariant item calibrations mean that item properties are unbiased by the population ability distribution. Because ability levels are included in the model, item parameter estimates are implicitly controlled for the abilities of the calibration sample. Very high ability or very low ability samples will yield the same item calibrations (see Hambleton et al., 1991).

Third, abilities have item-invariant meaning in item response theory. The PCCs show that the meaning of ability for performance applies to *any* calibrated item. Further, it can be readily shown that ability differences have invariant implications for item performance differences, regardless of the specific item (i.e., for the Rasch model). Andrich (1988) elaborated this principle. Fourth, item-referenced meaning is possible because ability level may be referenced directly to the items, as shown by the PCCs above. Diagnostic feedback may be given about an ability by direct reference to the items that persons at an ability level find hard or easy. Further, it is often revealing to examine the properties of items that fall at the threshold.

Fifth, handling missing data is a very significant practical advantage of item response theory. In fact, data may be so sparse that two persons do not even receive any common items. Missing data handling capabilities result from the item-invariant meaning of ability. Ability estimates are readily comparable over different item sets, such as from different test forms or from computerized adaptive testing. Because item parameters are included in the model, person ability level estimates are implicitly controlled for the properties of the items that were administered. Sixth, the standard errors of measurement reflect the appropriateness of the items for an individual. The information provided by an item for an ability depends mainly on probability of passing. The smallest standard error of measurement is obtained when many items are near a person's threshold level. Standard errors are also useful in selecting items for optimal precision for a person or a population.

Confirmatory IRT Models

A major advantage of IRT, as shown in the preceding section, is item-referenced meaning for ability. However, the exploratory IRT models do not elaborate the *nature* of the items that correspond to various ability levels. Most ability test items are complex problem-solving tasks that involve multiple processing stages. The unidimensional exploratory IRT model parameters will reflect a confounded composite of these influences, thus yielding parameters with unclear construct representation. Consequently, "enhancing" ability interpretations by showing representative items that correspond to the ability level will not be effective if the nature of the items cannot be clearly specified.

The confirmatory IRT models include parameters that can represent cognitive theory variables to describe the processing characteristics of the items. Two models will be described below.

Unidimensional Models

LINEAR LOGISTIC LATENT TRAIT MODEL. The linear logistic latent trait model (LLTM; Fischer, 1973) incorporates item stimulus features into the prediction of item success. For example, if the item stimulus features that influence processes are specified numerically (or categorically) for each item, as in a mathematical model of item accuracy, then the impact of processes in each item may be estimated directly as follows:

$$P(X_{ij} = 1 \mid \theta_j, \tau_k) = \frac{\exp(\theta_j - \Sigma_k \tau_m q_{ik})}{1 + \exp(\theta_j - \Sigma_k \tau_k q_{ik})} \qquad (19.3)$$

where q_{ik} is the value of stimulus feature k in item I, τ_k is the weight of stimulus feature k in item difficulty, and θ_j is the ability for person j.

To give an example, consider the matrix task in Figure 19.4. The Abstract Reasoning Test (ART; Embretson, 1995) contains matrix items that were designed to reflect Carpenter, Just, and Shell's (1990) theory of processing. Carpenter et al.'s (1990) theory emphasized working memory requirements and abstraction as underlying processing difficulty. Working memory requirements for solving a matrix problem were influenced by the number of relational tokens in a problem. Abstraction was influenced by abstract correspondence among elements (e.g., correspondence due to common properties rather than common objects) or null values.

ART items were generated mechanically by specifying 30 formal structures that defined various combinations of number and type of rules. In addition, some drawing principles were specified. The drawing principles specified whether objects in each position of the matrix were overlaid (objects placed inside other objects), fused (separate object appearing as a single object), or distorted (corresponding objects are perceptually distorted versions of each other). Five clone items that involved the same number and type of rules, but different stimuli and

attributes, were generated for each structure.

For the ART item in Figure 19.5, three relationships are involved. Two relationships are pairwise progressions in which the attributes are changing across the rows (orientation of the interior lines) or across the columns (i.e., number of lines). The third relationship is a distribution of three elements in which the oval, diamond, and rectangle are distributed to appear once in every row and column. None of the relationships involve a null value, and correspondence is not based on abstract properties. The drawing principles specified overlay (i.e., the lines are inside the shapes) but show no fusion or distortion.

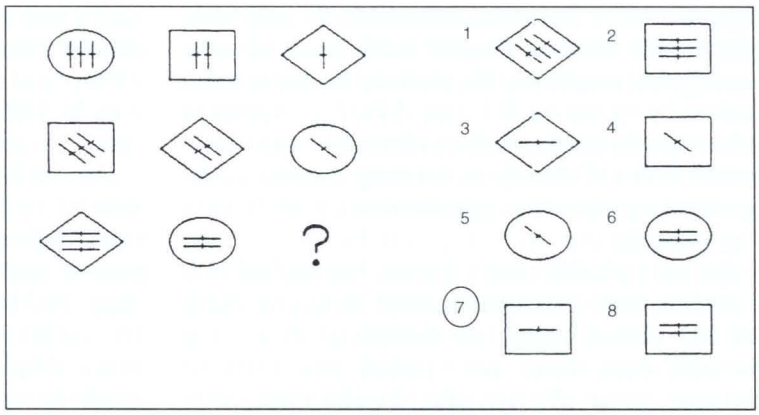


FIGURE 19.5. An item from the Abstract Reasoning Test.

To operationalize the Carpenter et al. (1990) theory into a mathematical model, each ART item was scored for number of rules and abstract correspondence, which operationalizes working-memory load and abstraction, respectively. Further, each item was also scored on the three drawing principles. Then, the 150 items were placed in 5 forms of 30 items each, with 4 additional items that appeared in every form to link estimates. The 5 forms were randomly assigned to 5 groups of about 250 participants each (see Embretson, 1995, for details).

Table 19.1 presents the LLTM estimates, standard errors, and t values that were obtained. Model fit for LLTM was comparable to a multiple correlation of .76. Although item difficulties are not fully explained by the variables in the model, moderately good prediction was obtained. Table 19.1 shows that the number of rules, which operationalizes working memory, is the most highly significant variable. However, abstract correspondence and two drawing principles were also significant.

An item can be decomposed into its processing contributions by multiplying the value of the stimulus feature q_{ik} times its weight. Thus, for the item in Figure 19.4, item difficulty is decomposed as follows:

$$\begin{aligned} &= .67(3) + 1.49(0) + .88(0) - .31(0) + .96(1) - 2.39. \\ &= 2.01 + 0 + 0 + 0 + .96 - 2.39 \\ &= .58. \end{aligned} \tag{19.4}$$

Thus, the item is predicted to be moderately difficult (.58), and the primary source of difficulty is working-memory load. Other items, in contrast, could be difficult primarily from other factors, such as abstract correspondence.

TREE-BASED REGRESSION. Sheehan (in press) has applied tree-based regression to enhance the meaning of the ability scale. In this method, clusters of homogeneous items with respect to skill components are located on the common IRT scale for item difficulty and ability. Persons' abilities can be described by the characteristics of the clusters at their

TABLE 19.1. LLTM Estimates for Extended Carpenter et al. Model on ART[a]

		LLTM		
Processing Variable	**Scored Variable**	**Weight τ_m**	**SE**	t
Working Memory Load	Number of Rules	.67	.09	7.14
Abstraction	Abstract Correspondence	1.49	.25	5.99
Drawing Principles	Distortion	.88	.29	3.05
	Fusion	−.31	.24	−1.31
	Overlay	.96	.20	4.73

[a] The model also included dummy variables to represent key position.

level. The item clusters are hierarchically organized; clusters are merged at higher levels based on common global properties. The cluster structure is determined by regressing IRT item difficulty on a set of binary predictors that reflect substantive item properties. As in LLTM analyses, items are scored on independent variables that represent sources of processing difficulty.

For example, Sheehan (in press) has applied tree-based regression to mathematical reasoning items. At the highest level, two clusters of items were formed; thus, items were parsed into two-class concrete versus abstract schemas. The binary variable that represented schema type had the strongest relationship to item difficulty. At the next level, the number of independent equations in the items defined the clusters. Number of independent equations had four levels, which were represented by three binary variables. These clusters for equations are nested within schema. Thus, within concrete schema, for example, items were parsed into clusters with one, two, three, or four equations. Similarly, items were also clustered by number of equations within abstract schema.

All clusters, regardless of hierarchical level, can be located on a common continuum for item difficulty and ability. Thus, for a given ability level, meaning may be enhanced by referring to the properties of the item clusters that correspond to that level.

SOME APPLICATIONS. The confirmatory IRT models described above have been applied to verbal and nonverbal items that measure intelligence. Embretson and Reise (in press) summarize these applications in a chapter on applications of IRT to cognition and life span. For example, applications of LLTM to nonverbal abilities includes studies of basic components in mathematical reasoning, abstract reasoning (i.e., matrix problems), geometric analogies, spatial visualization, and developmental balance problems. Applications of LLTM to verbal measures of ability include verbal comprehension, verbal analogies, and literacy. Applications of tree-based regression are just now appearing but include mathematical reasoning items and reading comprehension items.

IMPLICATIONS FOR MEASURING AND UNDERSTANDING INTELLIGENCE. Applications of both LLTM and tree-based regression can lead to (1) enhanced construct validity, (2) enhanced meaning for ability, and (3) item selection by cognitive properties. Each of these advantages will be elaborated, in turn.

For the first advantage, construct validity is enhanced by explicating the processes involved in item difficulty. The construct representation aspect of construct validity (see Embretson, 1983) is supported by elaborating the theoretical nature of the constructs reflected in test performance. It is most adequately studied by cognitive psychology methods such as mathematical modeling. The prediction model in LLTM and the item bundles in tree-based regression explicate the relative impact of various processes or knowledge structures on item difficulty.

For the second advantage, it should clearly be noted that the confirmatory unidimensional IRT models probably do not result in defining any new dimensions of intelligence. Instead, the IRT property of item-referenced meaning for ability is extended by describing the processing properties of items. Rather than merely show items that correspond to a person's ability, the investigator can show that items can be described by the processes and knowledge structures involved in their solution.

For the third advantage, decomposing items into cognitive components permits items to be selected by their cognitive properties. For example, if the construct to be measured is deemed to require a combination of processes, then these processes can be balanced across item subsets to reflect the desired combinations. For ART items, for example, if both abstraction and working memory are to be reflected in the measured ability, then items that include both sources of processing difficulty can be selected. This advantage can lead to selecting items that are more pure measures of targeted constructs.

Multidimensional Models

If ability test items are complex tasks with multiple processing stages, each stage may require a different ability. Multidimensional IRT models contain two or more abilities for each person. The confirmatory IRT models contain design structures to link items to underlying cognitive variables. The variables in the design structures can be derived

from cognitive psychology research to identify processes, strategies, or knowledge structures from item responses.

It should be noted that exploratory IRT models with multiple dimensions are also available. For example, Bock, Gibbons, and Muraki (1987) developed a multidimensional IRT model in which a linear combination of several abilities predicts item-solving probabilities. The exploratory multidimensional IRT models are very similar to factor analysis models. In fact, Bock et al. (1987) described their model as full-information factor analysis. It is full information in the sense that item responses are modeled, and hence the full dataset is used to estimate parameters. In contrast, factor-analysis attempts to model statistics that have been computed on the data; namely, correlations. The exploratory multidimensional IRT models are more appropriate than factor analysis for binary data or for data with discrete categories.

In this section, only confirmatory IRT models will be considered. Although the exploratory multidimensional IRT models are more justifiable for analyzing item level data, the models do not identify dimensions that differ much from factor analysis. Although confirmatory factor analysis models, in some applications, can identify new aspects of individual differences, little attention has been given to measurement from item level data. The confirmatory multidimensional IRT models, however, do have potential for identifying new aspects of individual differences.

In this section, only a few confirmatory IRT models will be presented. Confirmatory IRT modeling is a rapidly expanding area. This section cannot review all these developments, but several models will be mentioned here. The reader is referred to the Embretson (1983) *Handbook of Modern Item Response Theory* for more details on several models.

Models for Independent Processing Components

MLTM AND GLTM. The general component latent trait model (GLTM, Embretson, 1984) measures (1) abilities on covert processing components, (2) item difficulties on processing components, and (3) the relationship of item stimulus features to component processing difficulties. GLTM is a generalization of MLTM (Whitely,* 1980) and is a noncompensatory model appropriate for tasks that require correct outcomes on several processing components.

Two types of mathematical models as well as an IRT model are specified by GLTM. The first model in GLTM specifies how component outcomes are related to item solving. In GLTM, item success is assumed to require success on *all* underlying components. If any component is failed, then the item is not successfully solved. Thus, GLTM is a noncompensatory model that gives the probability of success for person j on item I, X_{ijT}, as the product of successes on the underlying components X_{ijm} as follows:

$$P(X_{ijT} = 1) = \Pi_m P(X_{ijm} = 1) \tag{19.5}$$

Although the component outcomes are covert, they may be operationalized by subtasks or by special constraints on the GLTM model without subtasks (see the following paragraphs for details).

The second mathematical model in GLTM specifies how item features influence component difficulty. Essentially, each component item difficulty is modeled by a linear combination of scored item features as in LLTM above. The third model in GLTM is an IRT model. A Rasch model gives the component success probabilities as a combination of ability and item difficulty. These three models are represented in GLTM in the following equation:

$$P(X_{ijT} = 1 | \Theta_j, b_i) = \Pi_m \frac{\exp(\theta_{jm} - \Sigma_k \tau_{km} q_{ikm})}{1 + \exp(\theta_{jm} - \Sigma_k \tau_{km} q_{ikm})}, \tag{19.6}$$

where τ_{km} is the weight of stimulus factor k in component m, q_{ikm} is the score of stimulus factor k on component m for item i, and θ_{jm} is the ability level of person j on component m. Like LLTM, item difficulty is replaced with a linear combination of the item features, which predict item difficulty.

To give an example, Maris (1995) applied MLTM to estimate two components, generation and evaluation, that were postulated to underlie success in synonym items. The results had several implications. First, Maris' results further elaborated the construct

* S. E. Embretson has published previously as S. E. Whitely.

representation of synonym items; the generation component was much stronger than the evaluation component in contributing to item solving. Second, generation ability and evaluation ability were measured for each person. Thus, diagnostic information about the source of performance may be obtained by comparing the relative strength of the two abilities. Third, the contributions of each component to the difficulty of each item could also be described. Such information is useful for selecting items to measure specified sources of difficulty.

Originally, both MLTM and GLTM required component responses, as well as the total item response, to estimate the component parameters. Now, however, GLTM can be applied to the total item task directly without subtasks if a strong cognitive model of item difficulty is available. That is, if stimulus features are known to predict item difficulty strongly, they may be used to place constraints on the GLTM solution when applied with Maris' (1995) missing-data algorithm. GLTM has been applied to measure individual differences in working-memory capacity versus control processing on abstract reasoning items (Embretson, 1995) and on spatial visualization items (Schmidt McCollam, 1998).

Models for Contrasting Experimental Task Conditions

Several IRT models can identify abilities by contrasting a person's performance over varying conditions. These IRT models interface well with contemporary cognitive experiments that use within-subject designs in which each person receives several conditions. In these experiments, construct impact is estimated by comparing performance across conditions. A similar approach can be applied in testing to measure individual differences in construct impact. Calculating performance differences directly for a person, say by subtracting one score from another, has well-known psychometric limitations (see Bereiter, 1963). Some new confirmatory IRT models, however, can provide psychometrically defensible alternatives. Several general models have been proposed, including Adams and Wilson (1996); DiBello, Stout, and Roussos (1995); and Embretson (1994, 1997). Like all structured IRT models, performance under a certain condition or occasion is postulated to depend on a specified combination of underlying abilities. Like confirmatory factor analysis, the specification is determined from theory.

These models may be illustrated by elaborating a special case, the multidimensional Rasch model for learning and change (MRMLC; Embretson, 1991), which measures a person's initial ability and one or more modifiabilities from repeated measurements. MRMLC contains a Wiener process design structure (shown below) to relate the items to initial ability and the modifiabilities. The Wiener process structure increases in the number of dimensions involved in performance across time or conditions. Complex cognitive data often have properties that correspond to the Wiener processing design structure, that is, increasing variances and decreasing correlations over time (see Embretson, 1991).

A general structured latent trait model (SLTM; Embretson, 1997) that includes MRMLC is given as follows:

$$P(X_{i(k)j} = 1 | \theta_j, b_i) = \frac{\exp(\Sigma_m \lambda_{i(k)m}\theta_{jm} - b_i)}{1 + \exp(\Sigma_m \lambda_{i(k)m}\theta_{jm} - b_i)}, \tag{19.7}$$

where θ_{j1} is initial ability level and $\theta_{j2}, \ldots, \theta_{jm}$ are modifiabilities between successive occasions or conditions, and b_i is difficulty for item i. The weight $\lambda_{i(k)m}$ is the weight of ability m in item i under occasion k. In MRMLC the weight is specified as 0 or 1, depending on the occasion. The Wiener process design structure determines which ability is involved on each occasion. For three occasions the structure, Λ, is specified as follows:

		Ability		
	Occasion	θ_1	θ_2	θ_3
	1	1	0	0
$\Lambda =$	2	1	1	0
	3	1	1	1

Thus, on the first occasion, only initial ability is involved. On the second occasion, both initial ability and the first modifiability are involved.

A recent application of MRMLC appears in Schmidt McCollam (1998). A dynamic test of spatial visualization ability with measures at three time points was analyzed by MRMLC for older and younger adults. Ability was estimated at time 1 (Pretest), after physical analogue training at time 2

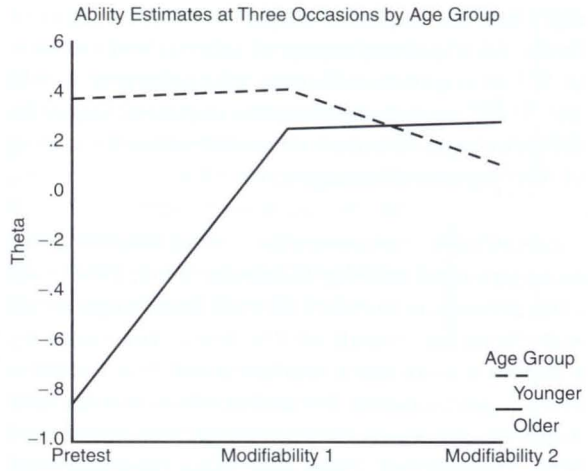


FIGURE 19.6. Means for initial spatial ability and two spatial modifiabilities in young and older adults.

(Modifiability 1), and after verbal-analytic training at time 3 (Modifiability 2) for both age groups. The Age x Occasion interaction was statistically significant, $F(2, 352) = 30.33$, $p < .0005$. It can be seen from Figure 19.6 that, although younger adults' initial pretest ability was greater than that of older adults, neither modifiability differed significantly. Analyses using raw scores failed to show the same trends. Because MRMLC is a Rasch-family model, it does have justifiable interval level measurement. Hence, MRMLC provides a more justifiable index of age group differences.

Equation 19.7 is obviously more general than MRMLC. Other design structures can be developed to represent specific comparisons or contrasts between conditions such as trends or Helmert contrasts and more.

Models for Distinct Classes of Persons

Several IRT models can identify groups of persons that differ qualitatively in their response patterns. On many cognitive tasks, persons have different knowledge bases or apply different strategies for item solving. The relative difficulty of the tasks depends on which knowledge structure or strategy is being applied. For example, in many spatial ability tests, items may be solved by either a spatial or a verbal strategy. Or, for another example, suppose that the source of knowledge for an achievement test differs between persons (e.g., formal education versus practical experience). In both cases, the groups are distinguished by a different pattern of item difficulties.

MIXED RASCH MODEL. The mixed population Rasch model (MIRA; Rost, 1990) has the following properties: (1) latent classes are identified from characteristic response patterns, (2) abilities are estimated for each person within each class, and (3) class membership is estimated for each person. In MIRA, IRT is combined with latent class analysis. The meaning of the ability level depends on the class because the item difficulties are ordered differently. The Mixed Rasch Model is given as follows:

$$P(X_{ij} = 1) = \sum_g \pi_g \frac{\exp(\theta_{jg} + b_{ig})}{1 + \exp(\theta_{jg} + b_{ig})} \qquad (19.8)$$

where $\theta_{jg} = \theta$ for person j in group g, b_{ig} = easiness for item I in group g, and π_g = the class size parameter or mixing proportion. Note that b_{ig} in this model is designated as item easiness. The b-value has a reverse sign from item difficulty. The constraint

$$\sum_g \pi_g = 1$$

designates the sum of the mixing proportions over all groups as equal to 1.

MIRA analyses enable estimation of class membership based upon different patterns of responses. Advantages of applying MIRA include (1) increased knowledge of construct representation for the test and (2) identification of a possible moderator variable (i.e., the latent class) that influences the prediction of criterion behaviors or other tasks.

An application that illustrates the properties of MIRA is a study by Schmidt McCollam (1998). A spatial visualization test was administered to a sample of adults and then analyzed with MIRA. Because it is well known that spatial solution strategies differ among persons, it was hypothesized that these strategy differences defined distinct latent classes. A goodness of fit test for MIRA indicated a two-class solution best fit the data. In one class, item difficulty patterns were highly related to spatial processing features of the items, whereas in the other it was not. Further, the external correlates of abilities from the two classes further indicated that the classes were spatial processing versus verbal-analytic processing.

An inspection of the mixing proportions indicated that 38% of the sample belonged to a verbal-analytic class and 62% belonged to a spatial processing class.

Thus, these results suggested qualitative as well as quantitative differences between persons. The results not only reflect widely held hypotheses about spatial processing but further allow assessment of individuals for processing strategies.

SALTUS: A DEVELOPMENTAL MODEL. Wilson (1989) designed the Saltus model as a developmental extension of the Rasch (1960) model to measure discontinuous stage changes in persons. Saltus is the Latin word for *leap*. The Saltus model measures state changes by using multiple tasks at each of various developmental levels. Wilson (1989) described the Saltus model in terms of Fischer et al.'s (1984) distinctions between first-order and second-order discontinuities. First-order discontinuities are sudden or abrupt changes in a single ability, whereas second-order discontinuities are these changes occurring in at least two domains. The Rasch model is used for first-order discontinuities, and the Saltus model, which estimates parameters for persons, domains, and levels, is used for second-order discontinuities.

The Saltus model also can be viewed as a refinement of Guttman's scalogram model (1944), which assumes that one item exists per level. A person is assigned to a given level based upon passing all previous items and failing all subsequent items. There are three major shortcomings of the scalogram model: (1) persons are discarded who do not adhere to the ordinal scalogram model, (2) the use of one item per level assumes that the exact nature determining the item's difficulty is known, and relatedly (3) the use of one item per level assumes no replication of task is needed. Wilson (1989) applied Rasch's probabilistic approach and interval scaling in Saltus to resolve the scalogram adherence problem. Further, Wilson (1989) noted that multiple tasks tied to cognitive theory can resolve the nature and replication problems.

A necessary part of first-order discontinuity for Wilson's Saltus model is segmentation, or the differences between difficulty across levels. Segmentation consistent with first-order discontinuity requires the item difficulties between levels be nonoverlapping. The Saltus model, in addition to θ_j and b_i, contains parameters to represent stages for persons and to represent the impact of stages on different types of items. An important aspect of Saltus is its extension of IRT to cognitive task data that otherwise would not fit IRT models. Further, the impact of stages on different types of tasks further elaborates the nature of developmental changes.

COGNITIVE DIAGNOSTIC ASSESSMENT. The rule-space methodology (Tatsuoka, 1983, 1984) classifies persons on the basis on their knowledge states, and measures overall ability level. For example, a person's score on a mathematical test indicates overall performance levels but does not usually diagnose processes or knowledge structures that need remediation. The rule-space methodology provides diagnostic information about the meaning of a person's response pattern. The meaningfulness of the diagnostic assessment depends directly on the quality of the cognitive theory behind the attributes and resulting knowledge states. The rule-space methodology has been applied to ability and achievement tests and to both verbal and nonverbal tests (see Embretson & Reise, in press, for a summary).

A basic rule space is defined by two dimensions, the ability level (namely θ_j from an IRT model), and by a fit index (ζ_j). Ability, of course, represents overall performance, and the fit index measures the typicality of the person's response pattern. For example, passing hard items and failing easy items is an atypical response pattern that would yield an extreme value for a fit index. Fit indices are calculated by comparing the person's response pattern to the predictions given by the IRT model.

Figure 19.7 plots both persons and knowledge states into the rule space from ability level and fit. Persons are classified into knowledge states by the distances of their response patterns from the locations of the knowledge states. Obviously, persons are plotted directly to the rule space because both ability level and fit are estimated. Locating a knowledge state requires some intermediate steps. First, an attribute incidence matrix is scored to reflect which attributes are required to solve each item. This is the first step in which cognitive theory is implemented. Second, knowledge states are defined from patterns of attributes. Knowledge states are extracted empirically by applying a Boolean clustering

algorithm to the attribute incidence matrix. Third, an ideal response pattern is generated for each knowledge state; that is, the ideal response pattern specifies the items that are passed and failed by someone in the knowledge state. Like a person, an ideal response pattern can be scored for ability (i.e., from total number of items passed) and for a fit index.

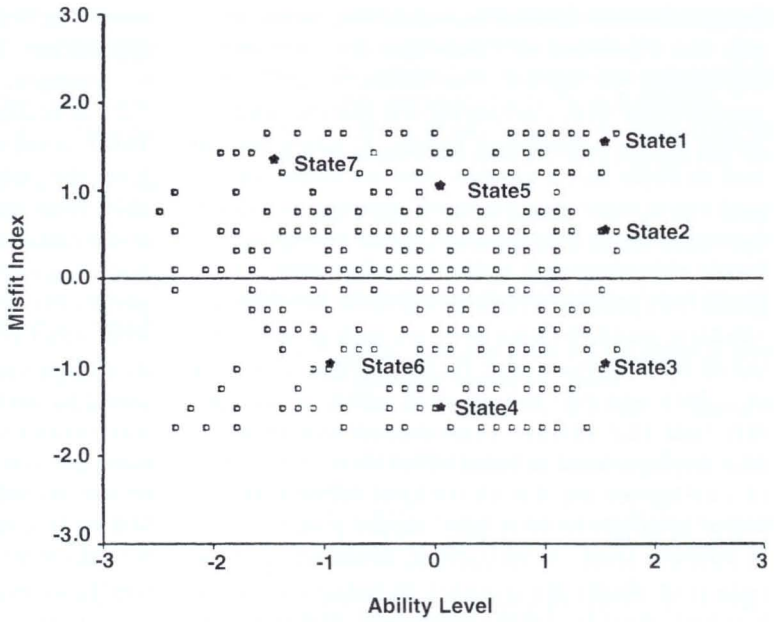


FIGURE 19.7. The rule space of ability level by misfit.

IMPLICATIONS FOR MEASURING AND UNDERSTANDING. Contemporary studies of intelligence have been strongly influenced by cognitive psychological theories. The confirmatory multidimensional IRT models can readily be interfaced with cognitive theory to understand intelligence. Unlike factor analysis, IRT focuses on the basis of item response rather than on decomposition of whole variables. The models include parameters to represent both persons and items.

Several advantages for measuring and understanding intelligence result from interfacing the confirmatory IRT models with cognitive theory. For items, the impact of various cognitive processes on item difficulty can be estimated from models such as LLTM or tree-based regression. In turn, these parameters can be used to select items by their processing demands. Recent research has also shown the potential to generate items from cognitive theory and to anticipate their psychometric properties by confirmatory IRT models (Embretson, in press). For persons, the confirmatory IRT models are more flexible in how abilities may be combined in processing an item. Multiple abilities are not restricted to a compensatory relationship as in factor analysis. In fact, compensatory models, such as linear combinations, do not interface well with contemporary theory in cognitive psychology. Noncompensatory IRT models better represent cognitive processes as independent events that must all be processed successfully.

Perhaps even more important is the potential of confirmatory IRT models to measure new aspects of individual differences. For example, qualitative differences in the basis of item responding can be assessed with the mixed Rasch model or with Tatsuoka's cognitive diagnostic assessment method. Further, individual differences in the component resources that underlie test performance can be assessed with the multicomponent latent trait model or the general latent trait model. Last, the differential sensitivity of persons to treatments or conditions that are varied during a testing session may be measured with the structured latent trait model.

These are just a few examples of the rapidly emerging field of confirmatory IRT models. Although such models are not fully implemented in testing as yet, relevant basic research to operationalize these models is expanding. At one major test publisher, Educational Testing Service, applications of diagnostic assessment and tree-based regression are now being considered.

SUMMARY

In this chapter we reviewed two types of psychometric models for impact on measuring and understanding intelligence. Factor analytic models and item response theory models are major psychometric approaches that are being applied to current measures of intelligence.

The factor analytic tradition was reviewed first because it has been the psychometric model with the

longest history. Until the mid-1970s, factor analysis was a predominant paradigm for understanding human intelligence. Understanding ability was synonymous with discovering the number and the nature of the underlying factors. Unidimensional and multiple factor models were reviewed for impact on current measurement practice. Although Spearman postulated a central factor in intelligence nearly a century ago, many contemporary intelligence tests provide an index of general intelligence. Similarly, multiple factor theories, such as those proposed by Thurstone and Guilford, have a history of at least one-half century and, again, contemporary tests that involve these factors remain available. Perhaps most influential for current measures of intelligence are the hierarchical factor theories. Many intelligence tests now report several scores at different levels of abstraction. Broad and narrow indices of ability are reported in tests such as the Stanford–Binet IV and the Differential Ability Scales, for example. Cattell and Horn's distinction between fluid and crystallized intelligence has been particularly influential, for many contemporary intelligence tests provide these estimates.

Item response theory methods were developed much more recently. For example, Rasch developed his IRT model in 1960. Currently, IRT is being applied in many contemporary tests to solve practical testing problems such as equating adaptive tests. The advantages of IRT were briefly reviewed, and two popular unidimensional IRT models were presented. Applications of these models have little potential to provide understanding of the nature of intelligence, however, or to define new aspects of individual differences.

However, a family of confirmatory IRT models is rapidly developing. These models have the potential to replace factor analysis as a tool for understanding intelligence. In fact, the exploratory multidimensional IRT models are equivalent to a full information factor analysis for item level data. In general, however, the IRT models are better interfaced with cognitive psychological approaches to understanding intelligence than are the factor analytic models for several reasons. First, it was shown that IRT models are applied to individual task responses like mathematical modeling in cognitive psychology. Second, the confirmatory IRT models utilize similar independent variables for modeling item difficulty as cognitive psychological approaches. The linear logistic latent trait model, for example, can utilize the same stimulus design data as mathematical modeling of response times. Third, confirmatory multidimensional IRT models have the potential to characterize qualitative variables more fully such as differing knowledge states and strategies. For example, the mixed Rasch model and diagnostic assessment can diagnose groups of person whose response patterns differ systematically from other response patterns. Fourth, confirmatory multidimensional IRT models have also been proposed to measure individual sensitivity to various interventions or conditions. Within-subject variations are commonly used in cognitive psychology to test hypotheses about processing mechanisms. Similar task variations can be applied in intelligence measurement to assess individual differences in sensitivity to processing.

If another *Handbook of Intelligence* is published several years from now, we predict that IRT-based approaches will predominate for measuring and understanding intelligence. Further, we anticipate a broad array of new types of individual differences to result from the application of the newer confirmatory IRT models. These new models can assess individual differences in knowledge structures, strategies, processing components, modifiability, and more if applied in the context of well-understood cognitive ability items.

REFERENCES

Adams, R. A., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory into practice*. Norwood, NJ: Ablex.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage Publications.

Arbuckle, J. L. (1995). Amos for Windows. Analysis of moment structures (Version 3.5). Chicago: SmallWaters.

Bennett, R. E., Seashore, H. G., & Wesman, A. G. (1984). *Differential aptitude tests: Technical supplement*. San Antonio, TX: Psychological Corporation.

Bentler, P. M. (1985). Theory and implementation of EQS: A structural equations program (manual for version 2.0). Los Angeles: BMDP Statistical Software.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. Harris (Ed.), *Problems in measuring change* (pp. 3–20). Madison, WI: University of Wisconsin Press.

Bock, R. D., Gibbons, R., & Muraki, E. J. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Burt, C. (1949). The structure of mind: A review of results of factor analysis. *British Journal of Educational Psychology, 19,* 100–111, 176–199.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review, 97,* 404–431.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1–22.

DiBello, L. V., Stout, W. F., & Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum Publishers.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Elliot, C. D. (1990). *Differential ability scales.* San Antonio, TX: Psychological Corporation.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika, 49,* 175–186.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–516.

Embretson, S. E. (1994a). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107–135). New York: Plenum Press.

Embretson, S. E. (1994b, April). *Structured multidimensional IRT models for measuring individual differences in learning or change.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Embretson, S. E. (1995). Working memory capacity versus general control processes in intelligence. *Intelligence, 20,* 169–189.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20,* 201–212.

Embretson, S. E. (1997). Structured ability models in tests designed from cognitive theory. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective Measurement III.* Norwood, NJ: Ablex, pp. 223–236.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380–396.

Embretson, S. E., & Reise, S. R. (in press). *Item response theory for psychologists.* Matawah, NJ: Erlbaum Publishers.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fischer, K. W., Pipp, S. L., & Bullock, D. (1984). Detecting discontinuities in development: Methods and measurement. In R. N. Emde & R. Harmon (Eds.), *Continuities and discontinuities in development.* Norwood, NJ: Ablex.

Guilford, J. P. (1948). Factor analysis in a test-development program. *Psychological Review, 55,* 79–84.

Guilford, J. P. (1959). Three faces of intellect. *American Psychologist, 14,* 469–479.

Guilford, J. P. (1967). *The nature of human intelligence.* NY: McGraw-Hill.

Guilford, J. P. (1977). The invariance problem in factor analysis. *Educational and Psychological Measurement, 37,* 11–19.

Guilford, J. P. (1985). A sixty-year perspective on psychological measurement. *Applied Psychological Measurement, 9,* 341–349.

Guttman, L. A. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139–150.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Holzinger, K. J., & Harman, H. H. (1941). *Factor analysis: A synthesis of factorial methods*: Chicago: University of Chicago Press.

Holzinger, K. J., & Swineford, F. (1939). A study in factor analysis: The stability of a bifactor solution. *Supplementary Educational Monographs, 48,* xi–91.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26,* 107–129.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24,* 417–441.

Humphreys, L. G. (1952). Individual differences. *Annual Review of Psychology, 3,* 131–150.

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications.* NY: Wiley (pp. 201–224).

Intelligence and its measurement: A symposium. (1921). *Journal of Educational Psychology, 12,* 123–147.

Joreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

Joreskog, K. G., & Sorbom, D. (1983). *LISREL VI: User's guide.* Uppsala: Department of Statistics.

Kaufman, A. S., & Kaufman, N. L. (1993). *Manual for the Kaufman Adolescent and Adult Intelligence Test (KAIT).* Circle Pines, MN: American Guidance Service.

Kelly, T. (1928). *Crossroads in the mind of man*. Stanford, CA: Stanford University Press.

Kelley, T. (1935). *Essential traits of mental life*. Cambridge, MA: Harvard University Press.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14,* 389–433.

Maris, E. M. (1995). Psychometric latent response models. *Psychometrika, 60,* 523–547.

Maxwell, S., & Delaney, H. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin, 97,* 85–93.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8,* 155–163.

Neale, M. C. (1994). Mx: Statistical modeling (2nd ed.). Richmond, VA: Author.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Raven, J. C. (1956). *Guide to using progressive matrices*. London: H. K. Lewis.

Raven, J. C., Court, J. H., & Raven, J. (1992). *Manual for Raven's Progressive Matrices and Vocabulary Scale*. San Antonio, TX: The Psychological Corporation.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271–282.

Schaie, K. W. (1985). *Schaie–Thurstone Adult Mental Abilities Test*. CA: Consulting Psychologists Press.

Schmidt McCollam, K. M. (1998). Latent trait and latent class models. In G. M. Marcoulides (Ed.), *Modern Methods for Business Research* (pp. 23–46). Hillsdale, NJ: Erlbaum.

Schmidt McCollam, K. M., & Embretson, S. E. (1998). Modifiability of spatial visualization performance in older and younger adults. Unpublished manuscript.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34,* 333–354.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Spearman, C. (1923). *The nature of intelligence and the principles of cognition*. London: Macmillan.

Spearman, C. (1927). *The abilities of man*. New York: Macmillan.

Spearman, C., & Jones, L. W. (1950). *Human ability: A continuation of "The abilities of man."* London: Macmillan.

Sternberg, R. J. (1977). *Intelligence, information-processing and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J., & Detterman, D. K. (Eds.). (1986). *What is intelligence? Contemporary viewpoints on its nature and definition*. Norwood, NJ: Ablex.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 34–38.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 94–110.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *The Stanford–Binet Intelligence Scale: Fourth Edition*. Chicago: Riverside.

Thurstone, L. L. (1938). *Primary mental abilities*. Psychometric Monographs, No. 1.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs*, No. 2.

Vernon, P. E. (1950). *The structure of human abilities*. New York: Wiley.

Vernon, P. E. (1961). *The structure of human abilities (2nd ed.)*. London: Methuen.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children-III*. San Antonio, TX: Psychological Corporation.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479–494.

Wilson, M. (1984). *A psychometric model of hierarchical development*. Unpublished doctoral dissertation, University of Chicago.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105,* 276–289.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson Tests of Cognitive Ability – Revised*. Chicago: Riverside.