Test Theory for a New Generation of Tests

Edited by
Norman Frederiksen
Robert J. Mislevy
Isaac I. Bejar
Educational Testing Service

LEA LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1993 Hillsdale, New Jersey Hove and London

12

Test Theory and the Behavioral Scaling of Test Performance

John B. Carroll
University of North Carolina at Chapel Hill

Behavioral scaling is proposed as a general term to cover various procedures for making test results directly interpretable in terms of what examinees know or can do. It is more inclusive than what has come to be known as *criterion-referencing*, which applies when tests are deliberately designed to provide behavioral information with reference to specific objectives of school learning. Test theory has an important role in behavioral scaling, but behavioral scaling requires use of a person characteristic function (PCF) rather than the item characteristic function. Problems that have arisen in efforts to scale tests behaviorally are discussed. Inasmuch as behavioral scaling is of particular importance and relevance in the case of cognitive ability tests, illustrations are given of behavioral scaling as applied to three subtests of the Woodcock–Johnson Psycho-Educational Battery—Revised, measuring *Gc*, *Gf*, and *Gv*, respectively. Data came from the norming versions of the tests given to 1,800 subjects ranging in age from 4 to 85. The behavioral scaling of these tests permits meaningful interpretation of differences over age groups.

INTRODUCTION

A perennial problem in psychological and educational measurement has been the interpretation of test results in terms of statements about what an examinee knows or can do, as opposed to statements about where the examinee stands relative to others. Conventional ways of interpreting test results have involved

some sort of numerical scale that is to be given what is essentially a normative interpretation. Scales for all kinds of tests have been developed; often they have tended to acquire meanings in their own right. For example, people tend to interpret points on the IQ scale or on the College Board Scholastic Aptitude Test scale in terms of subjectively judged degrees of intelligence or aptitude, but without clear ideas of what kinds of intellectual tasks individuals with given scores are able to perform.

Glaser (1963) introduced a distinction, now well recognized and widely used, between what he called *norm-referenced* and *criterion-referenced* tests. (He noted that such a distinction had been recognized earlier.) In his opinion, traditional test theory had led to the development of tests that might be quite satisfactory as norm-referenced tests for emphasizing differences among individuals, but that could be unsatisfactory for assessing differences between groups or the changes that might result from instruction. Subsequent developments in test theory have yielded some progress in developing domain- or criterion-referenced tests (Millman & Greene, 1989) and assessing their reliability (Feldt & Brennan, 1989, pp. 140–143). Nevertheless, even criterion-referencing of tests seldom addresses very well the problem of making clear what students' scores tell about what they can or cannot perform. There is also a tendency among test specialists to feel that the construction of criterion-referenced tests must follow different guidelines from those governing the construction of norm-referenced tests. Because the concept that underlies criterion-referencing ought to be applicable even to many types of norm-referenced tests, it appears that criterion-referencing is not the term of choice in referring in a general way to the process of relating test scores to behavior.

Instead, I propose the term *behavioral scaling*. This is not a new term. In 1955, I chaired an American Psychological Association symposium entitled "The behavioral scaling of psychological and educational tests." One of the participants was Ledyard Tucker, who reported what he called "some experiments in developing a behaviorally determined scale of vocabulary." I have at hand a copy of Tucker's (1955) paper, which from the standpoint of an early version of item response theory considered the problem of establishing a scale with units such that "an increase of one [unit] anywhere along the scale corresponds to an increase in probability of a correct answer from .5 to .7 for an item of appropriate difficulty and standard discriminating power." It was only in this sense that Tucker's scale would represent measurable differences in persons' "behavior." The nearest that Tucker came to alluding to what I would call behavioral scaling was in the fact that he offered specific multiple-choice vocabulary items to illustrate each of the 10 points of his scale, ranging, as he said, "approximately from the 15th percentile of fourth graders to the 85th percentile of [a] higher college group." For example, scale point 1 was illustrated by an item requiring recognition of the similar meanings of *throw* and *toss*, while scale point 10 was illustrated by an item concerned with the meaning of *imbue*. Presumably, by knowing an individual's score on this scale, one could predict what items the individual would probably get right, and the items that the individual would be likely to fail. However, Tucker did not consider prediction of the item difficulties themselves, or comment on the nature of the items found at given scale points.

The behavioral scaling of tests refers to the process of stating in behavioral terms what test results directly imply with regard to what examinees know or can perform. Behavioral scaling is thus one aspect of construct validity. It does not, or need not, extend to problems of external or predictive validity, although adequate behavioral scaling of tests could be of great assistance in dealing with problems of external validity.

In this chapter I propose to do three things:

1. Discuss the role of test theory in making behavioral scaling of tests possible or more meaningful.
2. Discuss a number of requirements and procedures in the behavioral scaling of tests.
3. Explore the use of certain procedures of behavioral scaling as applied to selected tests of cognitive ability.

THE ROLE OF TEST THEORY

Because behavioral scaling is most meaningfully applied to tests that are "good" or satisfactory in some general sense, test theory can help in the construction and analysis of such tests in terms of dimensionality, reliability, discrimination power, and range of difficulty level.

Behavioral scaling would be most readily applied to tests that approach unidimensionality. The construct intended to be measured should be essentially unidimensional in the sense that a single dimension can account for nearly all the variation in ability that is measured. Studies of the items measuring the National Assessment of Educational Progress (NAEP) Reading Proficiency Scale, for example, found that reading skill is essentially unidimensional (Zwick, 1987a, 1987b), despite the fact that logically, many separate skills can be identified in reading behavior. (For example, knowledge of any particular word or sentence structure could be regarded as a separate skill, but in point of fact, such skills tend to be highly correlated in any wide-ranging population.) Modern test theory has made considerable advances in procedures for assessing test dimensionality; Zwick's articles present an excellent review and application of such procedures. It is noteworthy that while reading comprehension as measured by NAEP procedures was found to be unidimensional, other curricular domains measured by NAEP, such as science (Mullis & Jenkins, 1988), appear to be

multidimensional to the extent that separate scales (though substantially correlated) have had to be established for reporting results.

At the same time, it is possible that behavioral scaling could be applied to tests that measure two or more constructs, that is, tests that would be found to be multidimensional. Results of using the linear logistic model of Fischer (1977, 1983), for example, could be behaviorally scaled if separate attention is devoted to each of the two or more constructs identified by such a model. The problem of multidimensionality also arises in connection with hierarchical models of intelligence, where a test could be found to measure not only a first-order factor but also a second-order factor, or even a third-order factor. It would be desirable to be able to apply behavioral scaling not only to the first-order factor but also to the higher-order factors.

Given that a test is found to be approximately unidimensional, an appropriate test theory model can be applied—for example, the Rasch one-parameter model, or a two- or three-parameter logistic model. (The three-parameter model is generally to be preferred because of its greater flexibility.) Such models ordinarily yield scales both for item (task) difficulty and for ability, and normally these are in the same metric. The item difficulty metric can take account of variations in item reliability, and procedures are provided for translating raw scores (or transforms such as proportion-correct) into the ability metric. Behavioral scaling can be applied either to the ability metric or to the item difficulty metric. As applied to the ability metric, statements can be made concerning the tasks that individuals at a given level on the ability metric are able to perform at some specified threshold level (e.g., with a probability of .50, or of .70, following Tucker's suggestion already mentioned, or of .80, following a suggestion of Bock, Mislevy, & Woodson, 1982). These statements would have to refer to behavioral scaling statements applied to the item difficulty metric—that is, statements concerning the nature of tasks placed at given levels on that metric.

It has been customary in item response theory to consider ability–difficulty relations in terms of the item characteristic curve (ICC)—that is, the function showing the increase of probability of correct response on a given item as ability increases. Such a function is useful in considering the operation of given items, but it fails to depict the overall functioning of a series of items in measuring an ability. For this purpose it seems more useful to use the person characteristic function (PCF), the function that shows the *decrease* of proportion correct, over items, as difficulty increases (Carroll, 1985). This function is based on the same mathematical expression as is the ICC, but variation in probability of correct responses is examined over item difficulty rather than over ability. Whereas for the ICC, each item can have a different ICC, for the PCF, each level of ability has a different PCF.

The PCF is necessarily based on aggregation of items; ideally, it would be assumed that all items have the same discrimination power. This assumption is in fact made by the one-parameter Rasch model; in the case of the two- or three-parameter logistic models such an assumption must be made in behavioral scaling.

The PCF has the further advantage over the ICC of depicting the variation over item difficulties in expected proportion correct for a given individual (or group of individuals with closely similar scores). If the item difficulties are behaviorally scaled, it is possible to describe an individual's or group's gradient of success in terms of statements about success at different described levels of difficulty. For example, if an individual were to be indexed as being at 200 on the NAEP Reading Proficiency Scale (implying, presumably, that the individual's success at that point is approximately 80%), a PCF function would also supply information as to expected success rates at 150, 250, or other points on the scale. Smith, Stenner, Horabin, and Smith (1989; see also Stenner, Horabin, Smith, & Smith, 1988) have proposed what they call a Lexile Scale of Reading Comprehension, based on reading difficulty measurements of texts or test items. They give illustrative PCF information for an individual with related lexile ability of 1000 as follows:

Text Difficulty	*Sample Titles*	*Predicted Success Rate*
600	(*Old Man and the Sea*—Hemingway)	96%
800	(*The Time Machine*—Wells)	90%
1000	(*Reader's Digest*)	75%
1200	(Encyclopedia)	50%
1400	(*The Washington Post*)	25%
1600	(*New England Journal of Medicine*)	10%

Here, the lexile ratings are linear transforms of Rasch-model item and ability indices; the behavioral scaling is in terms of typical reading material rated at specified levels of difficulty. Values on the lexile difficulty scale are pegged at a 75% success rate for individuals at given points on the scale.

PROCEDURES OF BEHAVIORAL SCALING

Can Behavioral Scaling Be Applied to Any Test?

The proper answer to this question is probably in the negative. A critical requirement is that the test be well constructed, be of adequate length and reliability, and consist of a series of items or tasks, of varying difficulties, that are more-or-less uniformly valid in measuring the construct intended to be measured. The role of test theory in test construction for behavioral scaling has already been discussed. The requirement of varying difficulties is made because

behavioral scaling needs reference points at different levels along a difficulty scale. If all items were of similar difficulties, behavioral scaling could refer only to a narrow band of the ability scale. Indeed, behavioral scaling depends on the assumption that an ability is defined in terms of individual differences in the points on a specified scale at which persons are able to perform at threshold levels.

It is often helpful, in producing behaviorally-scaled tests, to make an analysis of the behavioral domain that is to be sampled by the test. This process is illustrated in the "domain tests" constructed by Flanagan and associates (Flanagan et al., 1964, pp. 3-96) for Project TALENT. Domain tests were constructed in each of three areas: vocabulary, spelling, and reading comprehension. Work on the domain vocabulary test has apparently never been formally published, but from a draft manuscript in my possession (M. F. Shaycoft, 1968) it appears that the object was to develop a test that would indicate, for any given student, the absolute size of the student's English vocabulary in terms of number of word *meanings* the student knows. In developing this test, word meanings were systematically sampled from dictionary entries, and test items were drawn up in an effort to ascertain whether the respondent actually knew the specified word meaning, allowing for chance success by guessing in five-choice vocabulary items.

The domain test in spelling consisted of 150 words sampled systematically from the 5,000 words that were highest in frequency in the Thorndike-Lorge word list (Thorndike & Lorge, 1944). It was thus possible to make statements about the probability with which students could correctly spell words in this set of 5,000 words.

The domain for the tests in reading comprehension was defined in terms of the kinds of fiction and magazine nonfiction that students at the high-school level were likely to read.

Systematically defining the content domain is highly desirable for purposes of behavioral scaling, but it would not be necessary or even possible in all instances. Many well-constructed cognitive ability tests would be amenable to behavioral scaling, even though not based on systematic analyses of their content domains.

How Are Behavioral Scaling Statements to Be Framed?

The behavioral scaling statements that are to be attached to points on the item difficulty scale (or to points on the ability scale, with reference to thresholds of ability) can take any of several forms. Exactly what form or forms they may take can depend on the nature of the construct. For example, if the test is based on the sampling of a defined content domain, behavioral scaling can refer to attributes or categories of that domain.

A fourfold classification of types of behavioral scaling statements is proposed:

Type I: Specification of Illustrative Items, Tasks, or Relevant Materials

The simplest type is the specification of illustrative items, tasks, or materials associated with given scale points. This was the method used by Tucker (1955); he gave one illustrative item at each of 10 scale points, stating that an individual at that scale point would be expected to have a 70% chance of passing the item.

This type of behavioral scaling statement is often useful, particularly when it accompanies statements of other types. The main problem with it is that the test user cannot always be expected to have a clear idea about the difficulty of an item or what it presumably measures; furthermore, the behavioral scaling statement is not readily generalizable to other items of similar difficulty. The difficulty of a single item can be affected by many factors—the exact phrasing of the stem and the alternatives, for example. Even if more than one illustrative item is offered for a given scale point, as has been done for anchor points of the NAEP Reading Proficiency Scale (National Assessment of Educational Progress, 1985), it is sometimes left up to the test user or interpreter to form an impression of what behaviors the statement could be generalized to, or what level of ability or competence the scale point represents.

A variant of this is illustrated by the behavioral scaling statements attached to scores on the Project TALENT domain test of reading comprehension, namely, the specification of literary or magazine materials that a student with a given score would be expected to comprehend—a procedure also used by Smith et al. (1989) for their lexile scale of reading comprehension, as noted earlier.

Type II: Verbal Description of Competence Associated with Given Scale Points

A second type of behavioral scaling statement is the verbal description of a level of competence in terms of typical behaviors expected at that level. It is illustrated by the scaling statements attached to the various anchor points of the NAEP Reading Proficiency Scale (Beaton, 1987, pp. 381-390). These statements were drawn up by specialists in the teaching and testing of reading comprehension on the basis of groups of NAEP exercises selected to represent given anchor points.

It is difficult to formulate statements of this kind that are sufficiently meaningful, precise, and unambiguous. In many circumstances this form of description may be the most satisfactory that can be obtained, but the phraseology of such statements needs to be thoroughly thought through, edited, and checked.

A variant of this is illustrated by a procedure that was used to provide behavioral scaling of objective foreign language proficiency tests (Carroll, 1967). For each of several foreign languages, groups of students and teachers widely varying in proficiency were given a standard language proficiency interview and

assigned ratings on five-point scales of speaking and reading proficiency that have been widely accepted in the U.S. government for purposes of appraising language proficiency for foreign service. Each point on these scales is defined both in a brief description and with an amplified description; for example, point 3 on the speaking scale is described as "minimum professional proficiency," or in more detail, "able to speak the language with sufficient structural accuracy and vocabulary to satisfy representation requirements and handle professional discussions within a special field." Substantial correlations having been found between these ratings and objective proficiency test scores (r ranging from .63 to .82), the test scores were behaviorally scaled by equating them to the ratings and thus to the verbal descriptions provided for those ratings.

Type III: Use of Task or Content Parameters

A third general type of behavioral scaling uses parameters that describe the item or task content with reference to physical or other attributes (e.g., the difference in musical pitch, measured in hertz, in a pitch discrimination task) or with reference to a behavioral domain. Examples of the latter are the descriptions attached to scores on Project TALENT's domain tests in spelling—where the score indicates the number or percentage of words, in the first 5,000 words of the Thorndike-Lorge list, that the student can spell correctly. A further example is the lexile scale of reading comprehension (Smith et al., 1989) mentioned earlier, which is based on measures of vocabulary difficulty and sentence length in text material.

Such parametric descriptions have the virtue of clarity and objectivity, but test users may not always be able to grasp their meaning without thorough acquaintance with the characteristics of the task, scale, or domain. If, for example, one is told that a person is estimated to know 20,000 word meanings, one might be hard pressed to judge whether this signals a mediocre, average, or large vocabulary. Most test users would prefer to fall back on a normative interpretation. Despite its apparent scientific superiority, the parametric approach to behavioral scaling is beset with certain problems and requires that the test interpreter be adequately informed about the nature of the scale. But then, the problem is perhaps no more serious than that faced by a person familiar with the Fahrenheit scale of temperature who tries to learn the meaning of points on the Celsius scale.

Type IV: Reference to Levels of Cognitive Processing

A fourth type of behavioral scaling statement would make reference to the level of cognitive processing involved in correct response at a given item difficulty level. No illustrations of such statements have come to my attention, but I give such an illustration in connection with a cognitive ability test next.

BEHAVIORAL SCALING OF THREE COGNITIVE ABILITY TESTS

In order to explore procedures and problems in behaviorally scaling cognitive ability tests, a study was made of item data for selected subtests of the Woodcock-Johnson Psycho-Educational Battery—Revised (Woodcock & Johnson, 1989), or actually for the norming versions of those subtests. The data, kindly supplied by R. Woodcock (the chief author of the test), appeared to be particularly suitable for this purpose, for several reasons: (a) The tests are individually administered and require open-ended responses rather than choices among alternatives (as in typical paper-and-pencil tests); thus, the value of the chance success parameter, c, can be assumed to be zero. (b) Complete data were available on a large number of cases selected over a wide range of ages (4–85) and amounts of ability. (c) The test items were carefully devised, selected, revised, and arranged in order of difficulty by the test's authors to measure abilities over a wide range. (d) It is highly likely that each test is essentially unidimensional. (It was beyond the scope of this chapter to investigate the dimensionality of the tests.)

The objective was to illustrate how person characteristic functions can be derived and used, in conjunction with information on item difficulties and other characteristics of items, to assign substantive meanings to test scores in terms of increasing levels of cognitive performance. An effort was made to interpret these levels in terms of the knowledge bases and information-processing requirements necessary to attain them.

The cognitive ability portion of the Woodcock-Johnson Psycho-Educational Battery—Revised contains in all 21 subtests. For present purposes, data from three tests were selected for analysis:

1. *Picture Vocabulary.* According to the authors' manual, this test measures the ability to name familiar and unfamiliar pictured objects, and measures primarily verbal comprehension or crystallized intelligence (*Gc*). The norming version studied contained 33 items arranged in approximate order of difficulty. (The publication version contains 58 items.) Subjects are presented with these items, in order, in such a way as to determine the level of difficulty that each can attain. That is, a subject is given successive items until the subject can be confidently predicted to fail all or nearly all of the remaining items, after which testing is discontinued. Items are scored 1 for passing and 0 for failing; the total score is the number of passes.

2. *Concept Formation.* This test "measures the ability to identify rules for concepts when given both instances of the concept and non-instances of the concept. . . . This test primarily measures reasoning or fluid intelligence (*Gf*)." The revised test (identical to the norming version studied here) contains 35 items, administered in order of difficulty until the subject can be predicted to fail the

remainder. The score is the number of passes. The test is a controlled-learning task, and when subjects make errors in the learning phase, they are corrected. Part of the difficulty level of an item in this test is therefore a function of its location in the test.

3. *Visual Closure.* According to the authors, this test "measures the ability to identify a drawing or picture that is obscured in one of several ways. The picture may be distorted, having missing lines or areas, or have a superimposed pattern. This test primarily measures visual processing (*Gv*)." The norming version of the test contains 37 items arranged and administered in approximate order of difficulty. (The publication version contains 49 items.) The score is the number of passes.

Data on 1,800 individuals given these tests were processed by the program LOGIST (Wingersky, 1983) in order to determine estimates of item and ability parameters as a basis for behavioral scaling. This program uses maximum likelihood procedures to estimate the values of the IRT parameters a, b, and c for each item of a test and the values of θ for each individual in a sample given the test, on the basis of the expression for the expected probability of success on an item:

$$p = c + \frac{1 - c}{1 + \exp[-1.7a(\theta - b)]}, \qquad (1)$$

where a is the item discrimination parameter, b is the item difficulty parameter, and c is the pseudo-guessing parameter. In use of the program with the Woodcock-Johnson tests, the value of c was set equal to zero because there was little reason to believe that any responses arose from chance guessing. For each test, the analysis considered the 1,800 cases as a single sample despite great variation in age. For purposes of behavioral scaling, it was assumed that any given test score had the same meaning and corresponded to the same person characteristic function, regardless of the age of the subject.

The data were also analyzed by a program that determined empirical probabilities of passing selected groups of items for selected score groups.

BEHAVIORAL SCALING OF THE PICTURE VOCABULARY TEST

Performance on the Picture Vocabulary test involves two abilities: (a) ability to recognize a pictured object as something that has a name or commonly accepted appellation, and (b) ability to give that name from active recall vocabulary. In both cases, retrieval from long-term memory is required. Behavioral scaling thus should make reference to the parameters of the knowledge base (behavioral scaling Type III).

Word frequency in large counts of running words in English text is a variable that has often been used to estimate the familiarity of words in active or passive vocabularies, and hence it was considered as a possible basis for behavior-scaling the items. The total frequency of all acceptable responses for an item was assessed using the *American Heritage Word Frequency Book* (Carroll, Davies, & Richman, 1971). This was easily done for items with one or more single words as responses, but difficulties were encountered in the case of phrases like "movie house" and "panning gold." Also, because the frequency list does not include information about meanings, it was necessary to exclude frequencies of responses like *falls* (item 11) because *falls* could be a verb having little to do with *waterfalls* (another acceptable response). Thus, some judgment was required in assessing word frequencies. In the case of single-word responses (e.g., *padlock*, item 5) frequencies for related words—plurals, capitalizations, etc. (e.g., *PADLOCK*, *padlocks*, and *padlocked*)—were included. Frequencies for combined entries were computed by the method illustrated on p. 3 of the *Word Frequency Book* and stated in terms of the SFI (Standard Frequency Index) measure used there. The resulting values of SFI are listed in Table 12.1, which also presents various item statistics.

The behavior-scaling analysis was limited to 27 items (items 4–34, exclusive of 4 items considered invalid by the test's authors) because of missing data for the others. The correlation between SFI and b was disappointingly low in absolute magnitude, $-.528$, in contrast to correlations around $-.8$ found in other contexts (Carroll, 1980). The smallness of this correlation is probably due to the inaccuracy of the SFI measure in assessing the familiarity of the objects or activities pictured and the familiarity or recallability of the names. Because of the low correlation, it was decided to lay aside the word-frequency ratings and to use, instead, ratings of word familiarity found in Dale and O'Rourke's (1981) *Living Word Vocabulary*.

For each of some 44,000 English word meanings (often two or more meanings for a given word), the *Living Word Vocabulary* gives two ratings: (a) the grade level at which the word was tested (tests were given only at Grades 4, 6, 8, 10, 12, 13, and 16), and (b) the percent correct responses for students at that grade level (for the computations described below, the percentages were converted to proportions). Dale and O'Rourke's intent was to specify the lowest grade level at which the percent correct responses (in three-choice vocabulary items) would be at least 67%. In the present study, both ratings were tabulated for the most probable response word (and meaning) to each item in the Picture Vocabulary test. The multiple correlation of b (from LOGIST) with the grade rating and the logit of the Dale-O'Rourke proportion correct was computed; the resulting R was .858, both variables having significant contributions (for grade rating, $t = 5.84$, $p < .001$; for logit of proportion correct, $t = -2.66$, $p < .05$). The regression equation was

$$\text{Est}(b) = -1.1013 + .2719(\text{Grade}) - .7815[\text{logit}(P)].$$

TABLE 12.1
Picture Vocabulary Test: Item Statistics (N = 1,800)

Item Number	p	$\zeta(p)$	LOGIST Values		SFI	Grade Rating
			a	b		
4	.9851	−2.17	1.42	−2.69	52.6	3.95
5	.9687	−1.86	1.97	−2.09	38.0	3.06
6	.9578	−1.73	1.53	−2.09	51.7	1.35
7	.9627	−1.78	2.15	−1.97	51.0	2.83
9	.9285	−1.46	1.56	−1.79	46.0	3.44
11	.9040	−1.30	1.90	−1.53	49.2	3.44
12	.8800	−1.17	1.47	−1.48	56.2	4.11
10	.8572	−1.07	1.86	−1.28	44.4	5.61
14	.8278	−0.95	1.76	−1.16	39.1	3.27
15	.7814	−0.78	1.99	−0.96	53.8	5.41
19	.6806	−0.47	2.54	−0.56	52.0	5.93
16	.6734	−0.45	1.87	−0.55	33.6	6.23
17	.6555	−0.40	1.56	−0.51	55.0	5.72
18	.5361	−0.09	1.36	−0.12	47.4	6.99
20	.5376	−0.09	1.88	−0.11	47.9	6.36
21	.5216	−0.05	1.78	−0.06	38.3	9.48
23	.3114	0.49	1.82	0.62	40.5	7.52
25	.2966	0.53	2.18	0.65	41.7	8.62
22	.2695	0.61	2.18	0.74	45.2	8.96
26	.2210	0.77	2.36	0.90	23.8	11.74
27	.1784	0.92	2.36	1.06	44.8	7.39
28	.1429	1.07	2.22	1.22	37.3	9.35
31	.1301	1.13	2.16	1.29	35.6	14.30
30	.0839	1.38	2.51	1.50	37.3	11.29
32	.0760	1.43	2.77	1.53	38.1	14.44
34	.0474	1.67	1.43	2.05	42.7	10.29
33	.0339	1.83	1.60	2.14	41.3	7.92
Mean		−0.15	1.93	−0.19	43.9	7.00
SD		1.17	.37	1.38	7.5	3.39

Note. Items arranged in order of values of b. Items numbered 8, 13, 24, and 29 in the test were considered invalid by the test's authors; hence, data for these items were missing. Data for items 1–3 and 35–37 are not listed because these items were not used in the analysis, due to missing data either for LOGIST values or grade ratings, or both. Raw scores were based on only the 27 items listed.

For ready interpretation, the resulting values of Est(b) were rescaled linearly so that they had a mean (7.0) and standard deviation (3.388) identical to the mean and standard deviation of the grade ratings. The rescaled grade ratings for the items are listed in Table 12.1. For purposes of scaling them against values of b their regression on b was computed as

$$\text{Est(Grade rating)} = 7.4111 + 2.1125b.$$

The next task was to construct a graph showing estimated PCFs for selected raw scores as a function of item difficulties, with the behavioral scalings represented by the rescaled grade ratings aligned with the item difficulties. In constructing such a graph, it was necessary to consider the relation between raw scores (X) and the values of θ reported by the program LOGIST. In the general case this relation can be nonlinear, but in the present case it was clearly linear, with a correlation of .99916. The linear regression of θ on X was

$$\text{Est}(\theta) = -2.5976 + .1792X.$$

The values of Est(θ) were substituted into the three-parameter logistic equation 1, giving expected probability of success as a function of the difference between θ and b. Considering that c is assumed to equal zero, this equation becomes one for a two-parameter model:

$$p = \frac{1}{1 + \exp[-1.7a(\theta - b)]}. \qquad (2)$$

The value of a selected for use in this equation was the mean of the values of LOGIST a for the 27 items used in the analysis, namely, 1.93. (Experimentation with other possible values indicated that this value was likely to produce best fit to the empirical data.)

Fig. 12.1 is the resulting PCF graph showing a series of parallel lines giving, for selected raw scores or score intervals, expected probabilities of success as a function of item difficulty values b, shown on the baseline. Probabilities of success are given in terms of logits, where logit(p) = ln[p/(1 − p)], because this metric yields straight-line relations with b. Several horizontal lines are shown for logits of particular interest, namely, that for p = .5 (at logit = 0), the customary threshold level; that for p = .8 (at logit = 1.39), the mastery threshold suggested by Bock et al. (1982); and that for p = .90 (at logit = 2.20), a somewhat more stringent mastery threshold that I believe is useful to consider.

A further baseline is given in terms of the rescaled Dale–O'Rourke Grade Ratings, making it possible to predict what grade rating is likely to be attained for a given combination of raw score and probability. For example, for individuals in raw score group 16–17 (mean 16.5), 80% mastery can be predicted for words with grade ratings of 7.3, while 90% mastery can be predicted for words with grade ratings of 6.8. Placed along the grade-rating baseline are sample words drawn from the test, but the presumption would be that the statistics would apply to any words that might be sampled from the Dale–O'Rourke list. In this sense the Picture Vocabulary test can be considered to be behaviorally scaled.

Superimposed on the graph are also points derived from the empirical data for selected score groups and nine sets of items ordered in average value of b. At least by inspection, it can be seen that there is generally a close fit between the predictions of the model and the empirical data. The fit is especially good for logits lying between −3 and +3 (corresponding to probabilities between approximately .05 and .95); outside of these bounds, the fit is

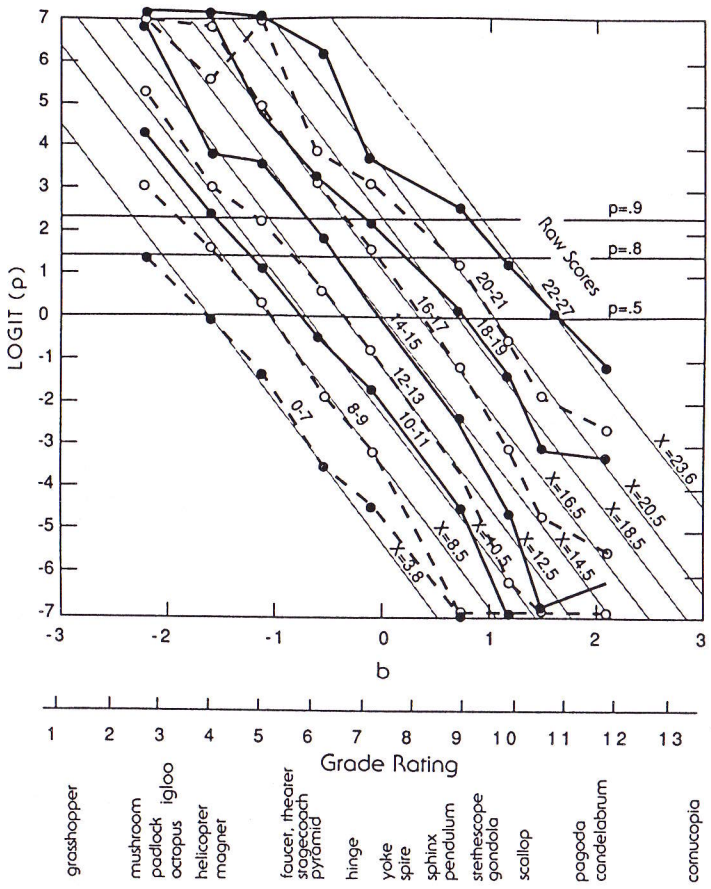

FIG. 12.1. Empirical PCF Curves for Picture Vocabulary Test data. Logit transforms of success probabilities are plotted against mean item difficulties (*b*) for 9 item sets, for each of 9 score-interval groups. Oblique parallel lines show expected logits for given score groups by the 2-parameter IRT model (Equation 2), with $a = 1.93$.

sometimes less good, mainly because of greater random error in probabilities near 0 or 1.

BEHAVIORAL SCALING OF THE CONCEPT FORMATION TEST

This appeared to be a test for which it might be possible to obtain a behavioral scaling criterion that measures the difficulty of the items as a function of information-processing characteristics (behavior-scaling Type IV). In this test, the examinee learns how to induce, from instances and noninstances that are

presented, rules by which stimuli are indicated as instances of a concept as opposed to noninstances. Noninstances and instances are presented as rows of stimuli, the noninstances at the left and the instances, each enclosed in squares, at the right. The test begins with a number of nonscored example items illustrating rules such as "[instance is] red" or "[instance is] two" (as opposed to there being only one entity). (Note that in this description I state the rules, in quotes, that are only implicit in the task, but in learning the task, examinees must be able to state such rules.) The first 15 items employ rules with only one term, the terms being selected from the oppositions red/yellow, round/square, big/little, and one/two that describe simple displays of geometric figures. In items 1–6, there is just one noninstance and one instance; in items 7–10, there are two noninstances and two instances; in items 11–15, four of each are given. Items 16–18 introduce two-term rules with the Boolean operator AND: for example, "[an instance must be] red AND square." In these items, there are six noninstances and two instances, a total of 8 stimuli. Items 19–21 introduce two-term rules with the operator OR: for example, "[an instance is either] yellow OR big." In these items there are two noninstances and six instances, a total of 8 stimuli. Items 22–24 introduce three-term OR rules, such as "[an instance must be] yellow OR square OR one." In these three-term items, there is one noninstance and there are seven instances.

From the start of the test up to item 24 the examinee is generally aware of the number of terms in the rule and whether it is an AND or an OR rule; further, wrong responses are corrected. From item 25 to the end at item 35, however, there is a mixture of item types, and wrong responses are not corrected. The examinee must induce, from the materials presented, what kind of rule is involved.

Given the structure of the task, the following characteristics of the items could be considered as elements in a behavioral-scaling criterion:

1. The number of noninstances (1–6).
2. The number of instances (1–7).
3. The total of (1) and (2) = the number of stimuli (2–8).
4. The number of terms (1–3).
5. Presence (1) or absence (0) of AND in the rule.
6. Presence (1) or absence (0) of OR in the rule.
7. Use (1) or nonuse (0) of the opposition *red/yellow*.
8. Use (1) or nonuse (0) of *round/square*.
9. Use (1) or nonuse (0) of *big/little*.
10. Use (1) or nonuse (0) of *one/two*.
11. Presence (1) or absence (0) of the item in the "mixture" part of the test (items 25–35).

The object was to use a linear combination of these variables in predicting item difficulty values, either in terms of LOGIST b or in terms of $\zeta(p)$, as shown in Table 12.2. It was found that predictions were significantly better against $\zeta(p)$, and therefore results are reported here only against this criterion. Because of the structure of the task and the use of dummy variables, however, the intercorrelation matrix of these variables (even excluding variable 3, which is the sum of variables 1 and 2) had such an amount of multicollinearity that it was singular. Investigation of the multiple regression of $\zeta(p)$ on these variables led to the decision to use only variables 3, 4, 5, 6, 7, 8, 10, and 11, whose intercorrelation matrix presented no singularity. The multiple correlation of these variables with $\zeta(p)$ was .9575, with variables 3, 5, 6, and 11 having t values significant with $p < .01$. The multiple correlation for variables 3, 5, 6, and 7 alone was .9464, all t values being significant with $p < .001$. The other variables were included in the final regression equation in order to provide further differentiation among items, and to take advantage of the slightly higher multiple R.

The final regression equation was

$$\begin{aligned}\text{Est}[\zeta(p)] = {} & -1.47 + 0.1000X(3) + 0.0306X(4) + 0.5696X(5) \\ & + 0.8324X(6) - 0.1465X(7) + 0.1288X(8) \\ & + 0.0314X(10) + 0.4237X(11). \qquad (3)\end{aligned}$$

Values of Est$[\zeta(p)]$ for the 35 items were rescaled to have a mean of 50 and a standard deviation of 10 and are called D-scores (analogous to T-scores). The D-score values for the items are listed in Table 12.2 and constitute readily interpretable measures of item difficulty based on item characteristics.

A PCF graph (Fig. 12.2) was constructed for this test by the same procedures that were used for the Picture Vocabulary Test. The regression of θ on raw scores X was sufficiently nonlinear to suggest that in computing Equation 2, values of θ be inserted corresponding to individual raw score values as given by a fitted curve, and this was done. (The linear correlation was .971.) The value of a was 1.61, the mean of the LOGIST values for individual items. For finding the correspondence between b and the difficulty scores D, the regression of D-scores on b was

$$\text{Est(D)} = 53.9159 + 8.1857b.$$

The graph provides a further baseline depicting the scale of D-scores.

As before, the graph also shows points from the empirical data, for selected item sets and score groups. There is an interesting tendency for the PCF curves to be flatter than expected for high values of logit(p) and the easier items. This is possibly due to the fact that the test involved a learning phase in which individuals were likely to make some errors on easy items even if they eventually

TABLE 12.2
Concept Formation Test: Item Statistics (N = 1,800)

			LOGIST Values		
Item Number	p	$\zeta(p)$	a	b	*D-Score*
5	.9282	−1.46	.83	−2.71	35.9
6	.9360	−1.52	1.12	−2.52	37.8
1	.9210	−1.41	.91	−2.49	33.7
2	.9229	−1.42	.95	−2.46	35.9
4	.8904	−1.23	.72	−2.38	37.8
3	.9061	−1.32	.89	−2.31	36.4
7	.8421	−1.00	1.29	−1.45	40.9
11	.8178	−0.91	1.44	−1.24	42.8
8	.7834	−0.78	.97	−1.20	38.9
9	.7851	−0.79	1.07	−1.16	39.4
15	.7966	−0.83	1.36	−1.13	42.8
10	.7912	−0.81	1.28	−1.12	40.9
12	.7842	−0.79	1.22	−1.10	45.0
14	.7611	−0.71	1.45	−0.91	47.0
13	.7200	−0.58	1.30	−0.73	45.5
16	.5806	−0.20	1.33	−0.15	51.9
25	.6020	−0.26	2.40	−0.13	42.3
26	.5779	−0.20	2.25	−0.06	47.3
33	.5138	−0.03	1.85	0.12	53.4
29	.4639	0.09	2.00	0.27	53.4
18	.4339	0.17	1.28	0.34	56.7
27	.4272	0.18	1.18	0.36	62.5
17	.4144	0.22	1.12	0.41	53.8
34	.4040	0.24	2.26	0.44	62.3
31	.3945	0.27	1.22	0.47	58.3
22	.3744	0.32	2.46	0.51	58.3
21	.3726	0.32	2.16	0.52	57.8
28	.3734	0.32	2.08	0.52	62.8
19	.3695	0.33	2.25	0.53	55.9
23	.3606	0.36	2.43	0.55	58.8
24	.3422	.0.41	2.00	0.60	58.8
20	.3168	0.48	2.20	0.67	60.5
35	.2995	0.53	2.63	0.70	65.2
32	.2850	0.57	2.30	0.75	65.2
30	.2612	0.64	2.26	0.82	64.7
Mean		−0.31	1.61	−0.48	50.0
SD		0.69	0.57	1.14	10.0

Note. Items in order of LOGIST b values.

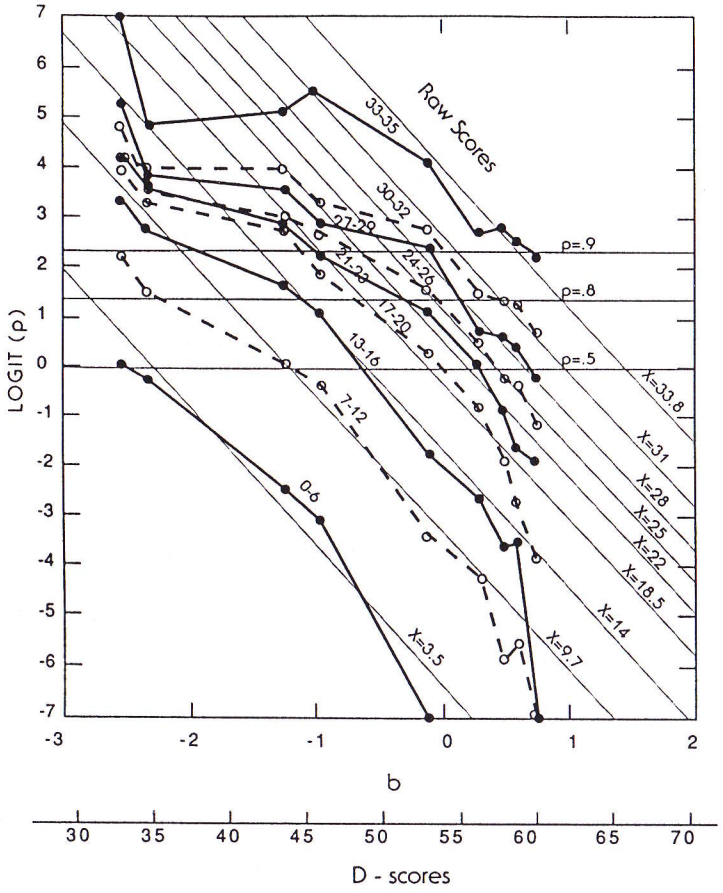

FIG. 12.2. Empirical PCF Curves for Concept Formation Test data. Logit transforms of success probabilities are plotted against mean item difficulties (*b*) for 9 item sets, for each of 9 score-interval groups. Oblique parallel lines show expected logits for given score groups by the 2-parameter IRT model (Equation 2), with $a = 1.61$.

were able to make high or at least above-average scores. Only a PCF analysis is able to display this phenomenon.

As an aid in interpreting the D-score difficulty values, Table 12.3 has been prepared to show what elements in the equation for predicting difficulty come into play for various items and difficulty values. The items are ordered in difficulty, and the actual values of the elements in Equation 3 are shown. As may be seen, for difficulty values (D-scores):

33–38. The examinee has to handle only two stimuli, with a one-term rule that does not involve AND or OR, and is not in the "mixture" part of the test.

TABLE 12.3
Concept Formation Test: Items Arranged in Order of Assigned Difficulty Values (D-Scores), with Elements Determining Them

		Variable							
Item Number	*D-Score*	*(3) Number Stimuli*	*(4) Number Terms*	*(5) AND*	*(6) OR*	*(7) Red/ Yellow*	*(8) Round/ Square*	*(10) One/ Two*	*(11) In Mixture Items*
1	33.7	2	1	0	0	1	0	0	0
2	35.9	2	1	0	0	0	0	0	0
5	35.9	2	1	0	0	0	0	0	0
3	36.4	2	1	0	0	0	0	1	0
4	37.8	2	1	0	0	0	1	0	0
6	37.8	2	1	0	0	0	1	0	0
8	38.9	4	1	0	0	0	0	0	0
9	39.4	4	1	0	0	0	0	1	0
7	40.9	4	1	0	0	0	1	0	0
10	40.9	4	1	0	0	0	1	0	0
25	42.3	2	1	0	0	0	0	0	1
11	42.8	8	1	0	0	1	0	0	0
15	42.8	8	1	0	0	1	0	0	0
12	45.0	8	1	0	0	0	0	0	0
13	45.5	8	1	0	0	0	0	1	0
14	47.0	8	1	0	0	0	1	0	0
26	47.3	4	1	0	0	0	1	0	1
16	51.9	8	2	1	0	1	0	0	0
29	53.4	8	1	0	0	0	1	0	1
33	53.4	8	1	0	0	0	1	0	1
17	53.8	8	2	1	0	1	1	0	0
19	55.9	8	2	0	1	1	0	0	0
18	56.7	8	2	1	0	0	1	1	0
21	57.8	8	2	0	1	1	1	0	0
22	58.3	8	3	0	1	1	1	0	0
31	58.3	8	2	1	0	1	0	0	1
23	58.8	8	3	0	1	1	1	0	0
24	58.8	8	3	0	1	1	1	1	0
20	60.5	8	2	0	1	0	1	1	0
34	62.3	8	2	0	1	1	0	0	1
27	62.5	8	2	1	0	0	1	0	1
28	62.8	8	2	0	1	1	0	1	1
30	64.7	8	3	0	1	1	1	0	1
32	65.2	8	3	0	1	1	1	1	1
35	65.2	8	3	0	1	1	1	1	1

39–41. The examinee deals with up to four stimuli, with a one-term rule that does not involve AND or OR, and is not in the "mixture" part of the test.

42–48. The examinee can deal with up to eight stimuli, but still with only a one-term rule that does not involve AND or OR, and is usually not in the "mixture" part of the test.

49–53. Like 42–48, but sometimes in the "mixture" part of the test.

54–57. Like 49–53, but the examinee can start to deal with two-term AND rules, sometimes with OR rules, but not in the "mixture" part of the test.

58–63. The examinee deals with two- and sometimes three-term rules using AND and OR, even in the "mixture" part of the test.

64–65. Like 58–63, but always involving three-term OR rules.

Item difficulty is clearly a function of the complexity of the tasks in terms of the number of stimuli to be inspected and dealt with, the number of terms in the rule, and the type of rule—particularly if it involves OR. The more difficult tasks call on rather complex processes of induction and deductive reasoning.

BEHAVIORAL SCALING OF THE VISUAL CLOSURE TEST

This test of what the authors call *visual processing* (*Gv*) requires examinees to name objects that are pictorially presented with varying degrees of indistinctness or obscuration. A behavioral scaling of this test could be of Type III in the sense that the scale would refer to attributes of the stimuli—that is, the degree of obscuration.

In most cases, it may be assumed that subjects, even young and inexperienced ones, would be able to name the objects if no obscuration were present. In many items, outline line drawings are presented with some or many parts of the lines omitted. Other items use one of five techniques to obscure the object: a superimposed horizontal grid of stripes (obscuring about 50% of the drawing); a two-way grid (obscuring about 67% of the drawing); a grid consisting of concentric circular bands; blurring by out-of-focus photography; and presenting the object in an unusual perspective. These techniques are not used in any systematic way; for example, there is only one item presenting an object in an unusual perspective, and photographic blurring is used only in two of the more difficult items. It is difficult to measure objectively the amount of obscuration; indeed, the item difficulty becomes the best operational measure of that, but it has an arbitrary unit of measure. Even if it were possible to measure the degree of obscuration, this might not predict difficulty very precisely because there is

an interaction between the way an object is obscured and how perceivable the object is. For example, in presenting a drawing of a diamond wedding ring with a superimposed circular grid, perceivability will vary depending on the relative positioning of the ring and the grid. The test is largely a product of the art and skill of the test authors in making pictured objects perceivable to different degrees.

Given the difficulty of objectively measuring the amount of obscuration and its effect, a rather simple and somewhat subjective procedure of behavioral scaling was employed, if only to illustrate what might be done for tests of this type. Items in which line drawings were obscured by omitting lines were judged (by the present author) for the amount (percentage) of lines deleted, on a scale that ranged from 5 to 95, and since these judgments had a correlation of .742 with $\zeta(p)$, the assigned criterion values were found according to the corresponding regression equation. Items using other techniques of obscuration were simply assigned the average of the $\zeta(p)$ values of those items. The resulting assigned difficulty values were rescaled to D-scores with a mean of 50 and a standard deviation of 10, and these D-scores are listed in Table 12.4.

A PCF graph (Fig. 12.3) was constructed for this test by the same procedures that were used for other tests. The scale of D-scores on the baseline is labeled with approximate percentages of obscuration at selected points. In constructing the graph, it was found that the regression of θ on raw scores X was highly linear, with a correlation of .9967; the regression of θ on X was

$$\text{Est}(\theta) = -5.5072 + 0.2463X.$$

In these computations, the data for W-J item 4 were excluded because of an abnormally high value of b, -12.17, for this extremely easy item. The value of a was 1.08, the mean value of LOGIST a for individual items. For finding correspondences between b and D-scores, the regression of D on b was

$$\text{Est(D)} = 53.2523 + 3.0319b.$$

DISCUSSION

Behavioral Scaling of Items

Behavioral scaling of items was most successful in the case of the Concept Formation test, where item difficulty predicted on the basis of objectively codable item characteristics had a correlation of .958 with obtained item difficulty. There was, of course, some capitalization on chance in the computations; it would be possible to cross-validate the predictions in a further study, particularly if item characteristics were more systematically varied. Behavioral scaling was largely successful in the case of the Picture Vocabulary test when appeal was made

TABLE 12.4
Visual Closure Test: Item Statistics (N = 1,800)

Item Number	p	$\zeta(p)$	LOGIST Values		Assigned D-Score
			a	b	
8	.9936	−2.49	0.64	−5.21	48.4
10	.9941	−2.52	0.72	−4.86	36.1
3	.9995	−3.29	1.46	−4.44	35.3
2	.9968	−2.72	1.14	−4.04	47.2
7	.9956	−2.62	1.07	−4.00	35.3
12	.9934	−2.48	0.95	−3.99	51.4
6	.9919	−2.40	0.94	−3.88	35.3
1	.9989	−3.05	1.83	−3.86	35.3
11	.9712	−1.90	0.68	−3.62	48.4
5	.9930	−2.46	1.66	−3.08	23.6
14	.9638	−1.80	0.83	−3.02	35.3
19	.9444	−1.59	1.01	−2.39	51.1
13	.9941	−2.52	1.31	−2.15	51.4
22	.8821	−1.19	0.89	−1.85	51.4
21	.8165	−0.90	0.71	−1.58	51.4
20	.8384	−0.99	0.99	−1.45	51.1
18	.8290	−0.95	1.05	−1.36	47.2
23	.6905	−0.50	1.18	−0.65	55.7
28	.6916	−0.50	1.47	−0.60	55.7
29	.6205	−0.31	1.50	−0.34	55.7
24	.6029	−0.26	1.25	−0.30	51.4
33	.4989	0.00	0.83	0.03	48.4
35	.3872	0.29	0.73	0.50	55.7
31	.3489	0.39	0.81	0.64	60.0
34	.3762	0.32	0.55	0.65	60.0
32	.2160	0.79	1.14	1.07	55.7
44	.1329	1.11	1.31	1.40	51.1
42	.1695	0.96	0.90	1.43	48.4
48	.0972	1.30	1.15	1.71	64.6
37	.1561	1.01	0.71	1.74	47.2
41	.0705	1.47	1.34	1.80	60.0
43	.0822	1.39	0.82	2.20	64.6
52	.0199	2.06	1.75	2.22	62.2
45	.0494	1.65	1.02	2.30	60.0
50	.0054	2.55	1.73	2.74	62.2
47	.0272	1.92	0.91	2.90	62.2
Mean		−0.56	1.08	−0.93	50.4
SD		1.68	.34	2.46	9.8

Note. Items in order of LOGIST b values. Items numbered 9, 15, 16, 17, 25, 26, 27, 30, 36, 38, 39, 40, 46, 49, and 51 in the test were considered invalid by the test's authors; data for these were missing. Data for item 4 were excluded because its value of p was .9995; LOGIST b = −12.17.

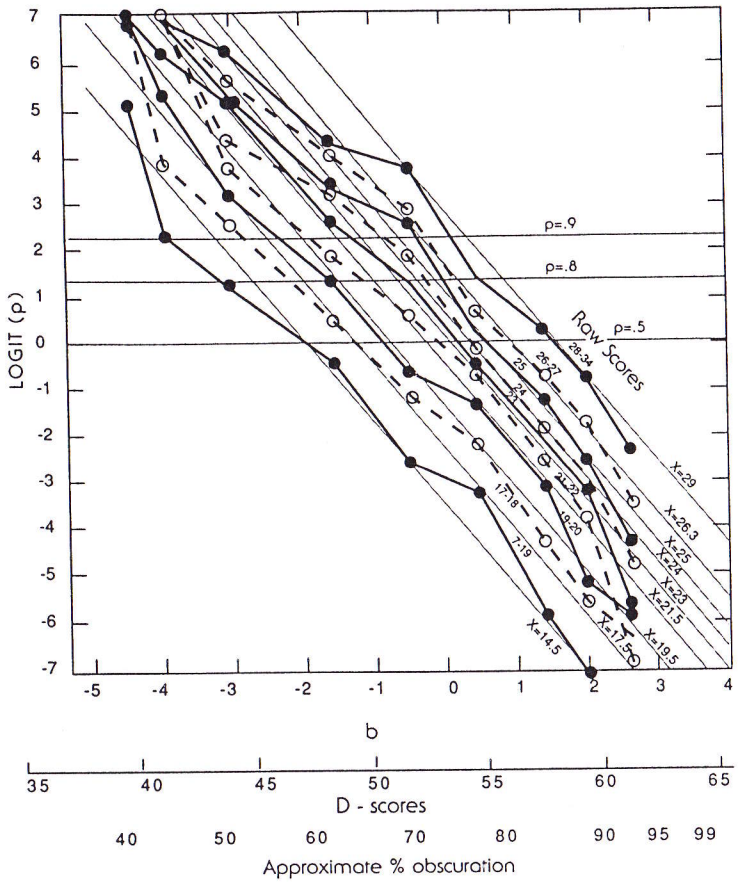

FIG. 12.3. Empirical PCF Curves for Visual Closure Test data. Logit transforms of success probabilities are plotted against mean item difficulties (b) for 9 item sets, for each of 9 score-interval groups. Oblique parallel lines show expected logits for given score groups by the 2-parameter IRT model (Equation 2), with a = 1.08.

to reasonably valid estimates of word familiarity. A correlation of .858 between predicted and obtained item difficulty was found. For the Visual Closure test, this correlation was only .793 and it was to some extent artifactual. It was limited by the fact that it was not possible to obtain objective estimates of the critical aspect of the items, namely, their degree of obscuration. It should be possible, however, to study the role of visual obscuration in a test of visual closure by a systematic variation of this variable.

For each of the three tests, the PCF (person characteristic function) procedure made possible the drawing up of graphs relating test scores to item difficulties. It is believed that this procedure permits a true behavioral scaling of test

scores in that it allows one to associate a given test score with a specifiable level of difficulty attained. For example, in the case of the Concept Formation test it is possible to interpret a given score as indicating a 90% chance of attaining a certain level of difficulty in concept formation, described in terms of the kinds of inductive inferences that the examinee is able to make.

Age Trends

Space limitations preclude presenting data on age trends in terms of score distributions for age groups. The behavior scaling of the test makes it possible, however, to interpret age trends in terms of specifiable levels of mastery attained on the average at different ages.

For example, on the Picture Vocabulary test the average child at age 5.5, with a raw score of 7.6 on 27 items, has a 50% chance of knowing words placed at grade 4.7, but a 90% chance of knowing words placed at grade 3.4 ("knowing" being defined in terms of the child's being able to recognize a picture and give an acceptable name for it). By age 19.5, the average person has a raw score of 19.2 on 27 items and a 50% chance of knowing words placed at grade 9.3, but a 90% probability of knowing words placed at grade 7.8.

Results from the Concept Formation test indicate that at age 5.5, the average child has a raw score of 9.6 and a 50% chance of performing tasks placed at a D-score of 42.1 (dealing with up to eight stimuli but only with simple one-term rules), but a 90% chance of performing tasks placed at a D-score of 32.8 (the bottom end of the range of tasks involving one-term rules and dealing with only two stimuli). By age 19.5, the average person has a raw score of 28.5 and a 50% chance of performing tasks at a D-score of about 59 (dealing with two- and sometimes three-term rules involving AND or OR), but a 90% chance of succeeding on tasks at a D-score of 53.6—dealing with one-term rules, sometimes in the "mixture" part of the test, and starting to deal with two-term rules involving AND, but not in the "mixture" part of the test where the examinee is "on his (or her) own."

Most of the age changes on the Visual Closure test occur between the ages of 4 and 16, at least for the average child. At age 5.5, the average child has a raw score of 17.5 and a 50% chance of perceiving objects with about 65% of the stimulus deleted, but a 90% chance of perceiving objects with less than 60% deleted. At age 19.5, the mean raw score is 25.1 and there is a 50% chance of perceiving objects with 80% deleted, but a 90% chance of perceiving objects with about 72% deleted. (These figures on percentages deleted must be taken with caution because they are not based on careful measurements. They are given only to suggest trends.)

It appears, in any case, that the technique of behavioral scaling presented here has great promise in clarifying the nature of different types of cognitive ability and describing given levels of performance. It would be interesting and informative to apply it to a wide range of cognitive ability and achievement tests.

ACKNOWLEDGMENTS

I am grateful to Dr. Richard W. Woodcock of Measurement Learning Consultants, Tolovana Park, OR, for supplying item response data from the norming of certain subtests of the *Woodcock-Johnson Psycho-Educational Battery—Revised*, to DLM Teaching Resources, Allen, TX, for permission to publish certain details about these tests, and to Dr. Albert E. Beaton of Educational Testing Service for arranging to have the item response data analyzed by the IRT program LOGIST.

REFERENCES

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983-84 technical report* (Report 15-TR-20). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Bock, R. D., Mislevy, R. J., & Woodson, C. E. (1982). The next stage in educational assessment. *Educational Researcher, 11*(3), 4–11, 16.

Carroll, J. B. (1967). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals, 1*, 131–151.

Carroll, J. B. (1980). Measurement of abilities constructs. In A. P. Maslow (Ed.), *Construct validity in psychological measurement: Proceedings of a Colloquium on Theory and Application in Education and Employment* (pp. 23–41). Princeton, NJ: Educational Testing Service.

Carroll, J. B. (1985). Defining abilities through the person characteristic function. In E. E. Roskam (Ed.), *Measurement and personality assessment* (pp. 121–131). Amsterdam: North Holland.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American heritage word frequency book*. Boston: Houghton Mifflin.

Dale, E., & O'Rourke, J. (1981). *The living word vocabulary: A national vocabulary inventory*. Chicago: World Book-Childcraft International.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education & Macmillan.

Fischer, G. R. (1977). Linear logistic test models: Theory and application. In H. Spada & W. F. Kempf (Eds.), *Structural models of thinking and learning* (pp. 203–225). Bern, Switzerland: Huber.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3–26.

Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycoft, M. F., Orr, D. B., Goldberg, I., & Neyman, C. A., Jr. (1964). *The American high school student*. Pittsburgh: University of Pittsburgh.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18*, 519–521.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). New York: American Council on Education & Macmillan.

Mullis, I. V. S., & Jenkins, L. B. (1988). *The science report card: Elements of risk and recovery. Trends and achievement based on the 1986 National Assessment*. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1985). *The reading report card: Progress toward excellence in our schools; Trends in reading over four national assessments, 1971–1984*. Princeton, NJ: Educational Testing Service.

Shaycoft, M. F. (1968). *Development and preliminary analysis of a test to estimate size of vocabulary*. Palo Alto, CA: American Institutes for Research.

Smith, D. R., Stenner, A. J., Horabin, I., & Smith, M. (1989, May). *The lexile scale in theory and practice: Final report for NIH Grant HD-19448*. Paper presented at the meeting of the International Reading Association, New Orleans.

Stenner, A. J., Horabin, I., Smith, D. R., & Smith, M. (1988). Most comprehension tests do measure reading comprehension: A response to McLean and Goldstein. *Phi Delta Kappan, 69*, 765–767.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30000 words*. New York: Bureau of Publications, Teachers College, Columbia University.

Tucker, L. R. (1955, September). *Some experiments in developing a behaviorally determined scale of vocabulary*. Paper presented at the meeting of the American Psychological Association, San Francisco.

Wingersky, M. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 45–56). Vancouver, BC: Educational Research Institute of British Columbia.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson Psycho-Educational Battery—Revised: Tests of Cognitive Ability*. Allen, TX: DLM Teaching Resources.

Zwick, R. (1987a). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24*, 293–308.

Zwick, R. (1987b). Assessment of the dimensionality of NAEP Year 15 reading data. In A. E. Beaton, *Implementing the new design: The NAEP 1983–84 technical report* (Report 15-TR-20, pp. 245–284). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

13

A Generative Approach to Psychological and Educational Measurement

Isaac I. Bejar
Educational Testing Service

INTRODUCTION

Response generative modeling (RGM) is an approach to psychological measurement that involves a "grammar" capable of assigning a psychometric description to every item in a universe of items and is also capable of generating all the items in that universe (Bejar & Yocom, 1991). Such an approach to measurement, if feasible, could have at least three important implications. First, the interpretation of scores from a generative instrument would be greatly facilitated because the process for generating the item is explicitly stated. Second, the possibility of generative modeling implies that we have a complete understanding of the underlying response process. Such knowledge might allow us, in turn, to abandon the multiple-choice format in favor of open-ended formats, a long-standing desire of psychometricians (e.g., Frederiksen, 1990), but without the expense associated with scoring open-ended responses. In other words, the same knowledge base that is used to create items can be brought to bear on the scoring of open-ended responses. Third, the ability to assign a psychometric description to an item is the key ingredient in what might be called *intelligent test development aids*. Job aids, in general, are rapidly becoming the key to increased productivity in many fields (e.g., Harmon, 1986). In a testing context, test development job aids might become essential if bills to outlaw pretesting succeed in becoming law (because it is through pretesting that test developers estimate the difficulty of an item before the test is administered in a final form), especially in light of growing statistical theory designed to allow equating tests