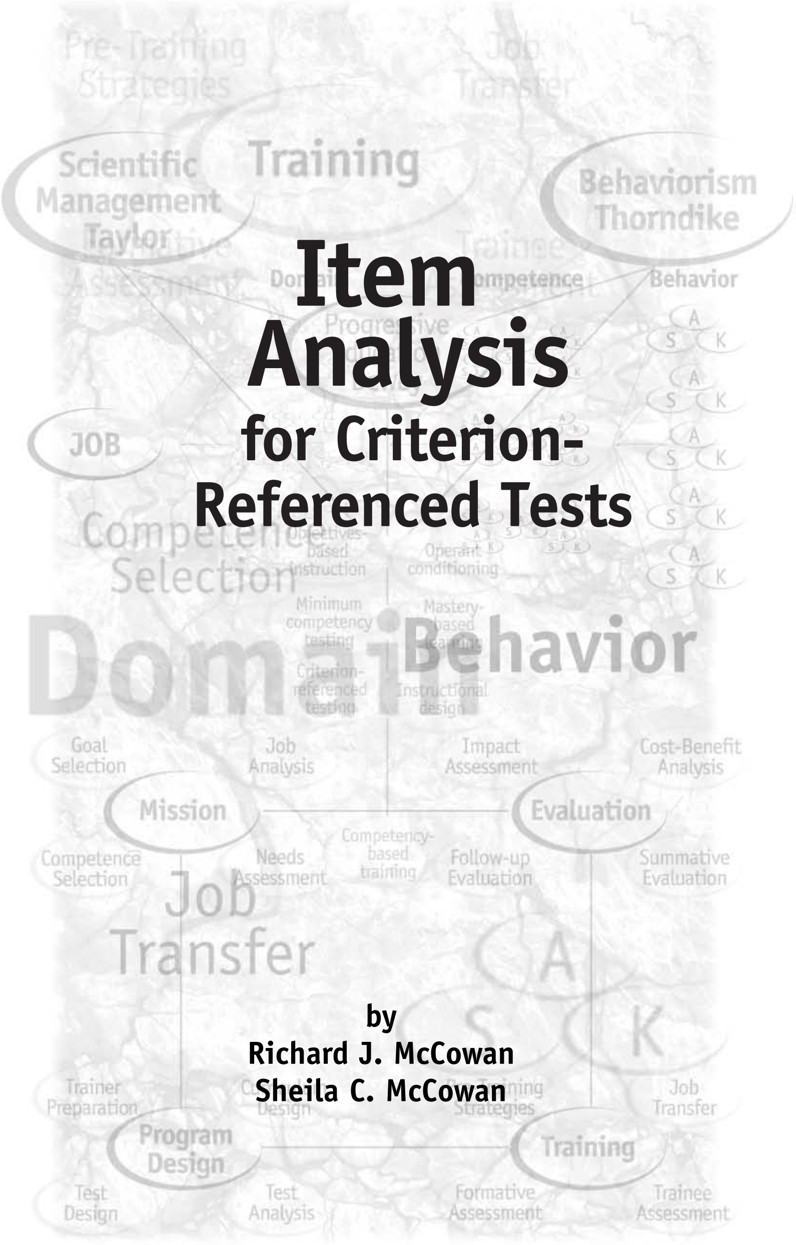
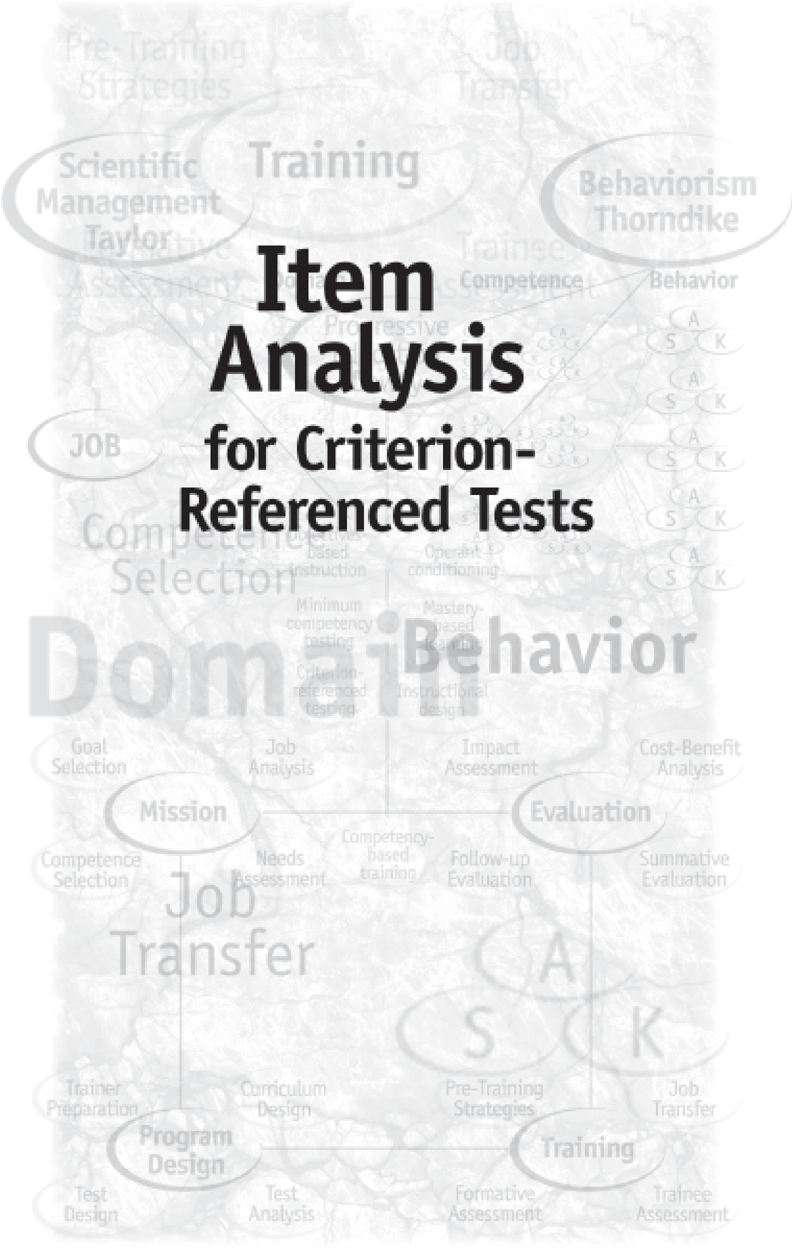
Item Analysis for Criterion-Referenced Tests

by
Richard J. McCowan
Sheila C. McCowan



Item Analysis for Criterion-Referenced Tests



CDHS
Center for Development
of Human Services

The Center for Development of Human Services is a continuing education enterprise of the Research Foundation of the State University of New York and a unit of the Graduate Studies and Research Division at Buffalo State College (SUNY).
Funding is provided by the New York State Office of Children and Family Services.

Center for Development of Human Services
Robert N. Spaner - Chief Administrative Officer
bobs@bsc-cdhs.org
Richard J. McCowan, Ph.D. - Director, Research & Evaluation Group
dickm@bsc-cdhs.org
Buffalo State College (SUNY)
1695 Elmwood Avenue
Buffalo, New York 14207-2407
Tel.: 716.876.7600
Fax: 716.796.2201
http://www.bsc-cdhs.org

Introduction

Item analysis uses statistics and expert judgment to evaluate tests based on the quality of individual items, item sets, and entire sets of items, as well as the relationship of each item to other items. It "investigates the performance of items considered individually either in relation to some external criterion or in relation to the remaining items on the test" (Thompson & Levitov, 1985, p. 163). It uses this information to improve item and test quality. Item analysis concepts are similar for norm-referenced and criterion-referenced tests, but they differ in specific, significant ways.

With criterion-referenced tests, use norm-referenced statistics for pretest data and criterion-referenced statistics for posttest data. This suggestion assumes that untrained persons will know relatively little about pretest material, so the assumptions on which norm-referenced statistics are based are applicable. Once people are trained, a test is criterion-referenced, and criterion-referenced statistics must be used.

Validity

Validity is the extent to which a test measures what it is supposed to measure. It is the most critical dimension of test development. Simply stated, validity is what a test measures and how well it does this (Anastasi, 1954; Anastasi & Urbani, 1997). Validity is a crucial consideration in evaluating tests. Since new commercial tests cannot be published without validation studies, it is reasonable to expect similar evidence of validity for tests that screen individuals for high stake decisions such as promotion, graduation, or certification.

With minor modifications, Cronbach's (1949) concept of validity has remained consistent over the last 50 years. Cronbach (1949, p. 48) said that validity was the extent to which a test measures what it purports to measure and that a test is valid to the degree that what it measures or predicts is known. He identified two basic categories of validity including logical and empirical. Logical validity is a set of loosely organized, broadly defined approaches based on content analysis that includes examination of operational issues and test-taking processes. Content validation requires that test makers study a test to determine what the test scores truly mean.

In 1954 the American Psychological Association (APA) defined four categories of validity including content, predictive, concurrent, and construct. In 1966, the association combined predictive and concurrent validity into a single grouping called criterion validity (American Psychological Association, 1966) which remains the current classification (American Educational Research Association, American Psychological Association, & National Council on Measurement and Education, 1985). These aspects of validity are often mistakenly considered as three types of validity rather than a concept about how a score can be interpreted.

Types of Validity

Face validity estimates whether a test measures what it claims to measure. It is the extent to which a test seems relevant, important, and interesting. It is the least rigorous measure of validity.

Content validity is the degree to which a test matches a curriculum and accurately measures the specific training objectives on which a program is based. Typically it uses expert judgment of qualified experts to determine if a test is accurate, appropriate, and fair.

Criterion-related validity measures how well a test compares with an external criterion. It includes:

> *Predictive validity* is the correlation between a predictor and a criterion obtained at a later time (e.g., test score on a specific competence and caseworker performance of a job-related tasks).
>
> *Concurrent validity* is the correlation between a predictor and a criterion at the same point in time (e.g., performance on a cognitive test related to training and scores on a Civil Service examination).

> *Construct* validity is the extent to which a test measures a theoretical construct (e.g., a researcher examines a personality test to determine if the personality typologies account for actual results).

In *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement and Education, 1985) stated:

> Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is a process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself. (p. 9)

They noted that professional judgment guides decisions about forms of evidence that are necessary and feasible regarding potential uses of test scores.

In 1955 Cronbach and Meehl amplified the concept of construct validity by introducing the concept of a nomological net. This net included the interrelated laws that support a construct. In 1971, Cronbach said that “Narrowly considered, validation is the process of examining the accuracy of a specific prediction or inference made from a test score” (p. 443). In 1989 Cronbach moderated this concept by acknowledging that it was impossible to attain the level of proof demanded in the harder sciences with most social sciences constructs.

A concept is an abstraction formed by generalizing from particulars, while a construct is a concept deliberately invented for a specific scientific purpose (Kerlinger, p. 28). The constructs on which a test is based relate specifically to the domain of competencies that are tested by items included on the test. Construct validity is to the extent to which a test is based on relevant theory and research related to a defined domain of behavior.

Cureton (1951 provided a definition similar to Cronbach when he noted that the essential question of test validity was how well a test did what it was employed to do. Validity, therefore, was the correlation between an actual test score and the "true" criterion score. By the early 1950's, other types of validity had been identified (e.g., factorial, intrinsic, empirical, logical) (Anastasi, 1954). Messick (1989) expanded the definition by stating "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 18).

Empirical validity emphasized factor analysis based on correlations between test scores and criterion measures (Anastasi, 1950). However, test makers must interpret correlational studies cautiously because spurious correlations may be misleading (e.g., high positive correlations between children's foot size and reading achievement).

In 1957 Campbell introduced the notion of falsification in the validation process due to spurious correlations, and he discussed the importance of testing plausible, rival hypotheses. Campbell and Fiske (1959) expanded this concept by introducing the multitrait-multimethod approach and convergent and divergent (or discriminant) validity.

Recently, Messick (1989) discussed the importance of considering the consequences of test use in drawing inferences about validity and added the term *consequential validity* to this list. He noted:

> Validity is an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores. As such validity is an inductive summary of both the adequacy of existing evidence for and the appropriateness of potential consequences of test interpretation and use (Messick, 1988, pp. 33-34).

Improving Test Validity

Anastasi (1986) described validation as a process built into the tests during planning and development.

> Validity is thus built into the test from the outset rather than being limited to the last stages of test development. . . . the validation process begins with the formulation of detailed trait or construct definitions derived from psychological theory, prior research, or systematic observation and analyses of the relevant behavior domain. Test items are then prepared to fit the construct definitions. Empirical item analyses follow with the selection of the most effective (i.e., valid) items from the initial item pools. Other appropriate internal analyses may then be carried out, including factor analyses of item clusters of subtests. The final stage includes validation and cross-validation of various scores and interpretive combinations of scores through statistical analyses against external, real-life criteria. (p. 3)

Many tests are flawed because they focus on insignificant or unrelated information, disproportionately test one segment of curriculum and ignore other sections, or include poorly written, confusing test items.

The following test development procedures will increase test validity:

Specify the instructional objectives.

Describe the behaviors that will be tested.

Describe the conditions under which the test will be given.

Determine number of items required.

Prepare domain specifications.

Determine testing time required.

Select item formats.

Write test items matched to specific objectives or sets of objectives.

Use a number of items that proportionately reflects the amount of training time devoted to objectives.

Establish a vocabulary level appropriate for the intended trainees.

Have experts independently review the items.

Prepare clear, simple test instructions.

Establish standards for mastery.

Use a specification table to match topics, objectives, and skill levels which trainees are expected to attain for each item.

Estimate what proportion of the curriculum addresses each competence.

Prepare scoring keys.

Prepare report and table formats.

Administer the draft test to a group comparable to people who will be trained.

Interview examinees to determine if they felt the test was appropriate.

Identify “poor” items such as those answered incorrectly by many examinees.

Score items (0,1) for each trainee in the instructed and uninstructed groups.

Compute a difficulty index for each item for instructed and uninstructed groups.

Compute the discrimination index for each item.

Summarize item statistics for each item.

Evaluate how items discriminate between masters and non-masters for each objective.

Analyze choice of distracters for questionable multiple-choice items

Revise test based on discussion with criterion group and item statistics.

Assemble items into the final test.

Reliability

In 1904, Spearman described true-score-and-error model which was accepted as “classical” reliability theory for the next 50 years. Classic reliability is a test score that includes a true score and random error. A true score is a theoretical, dependable measure of a person’s obtained score uninfluenced by chance events or conditions. It is the average of identical tests administered repeatedly without limit. Identical implies that a person is not affected by the testing procedure, which is unlikely to occur. A raw score, which is the number of points a person obtains on a test, is the best estimate of the true score. Chance conditions, such as test quality and examinee motivation, may underestimate or overestimate true scores (Thorndike, 1982).

In 1953, Lindquist described a multifaceted reliability model that was adapted by Cronbach, Gleser, Nanda, and Rajaranam (1972). More recently, reliability has been expanded to include generalizability theory, domain mastery, and criterion-referenced testing.

Reliability, which is the best single measure of test accuracy, is the extent to which test results are consistent, stable, and free of error variance. It is the extent to which a test provides the same ranking of examinees when it is re-administered and is measured by Coefficient Alpha or KR-20. A reliable test may not be valid. A yardstick only 35 inches long will measure consistently, but inaccurately, resulting in invalid data. Reliabilities as low as .50 are satisfactory for short tests of 10 to 15 items, but tests with more than 50 items should have reliabilities of .80 or higher. If reliability is less than .80, a single test score should not be used to make important decisions about individuals. Low reliability or error variance results from chance differences and is affected by different factors such as:

> Variations in examinee responses due to physiological or psychological conditions such as amount of sleep or motivation.
>
> Too many very easy or hard items.
>
> Poorly written or confusing items.
>
> Items that do not test a clearly defined, unified body of content.
>
> Changes in curriculum not reflected in the test.
>
> Testing conditions such as temperature, noise, or apparatus functioning.
>
> Errors in recording or scoring.
>
> Test length (i.e., longer tests have higher reliability.

Lower score variance increase reliability.

Difficult items that cause excessive guessing reduces reliability.

Reliability coefficient is a generic term that refers to various type of reliability measures.

Types of Reliability Coefficients

Stability (test-retest) is the correlation between two successive measurements using the same test. The reliability may be spuriously high effect due to item recall if the time between test administrations is too brief or too low if too much time elapses between the pretest and posttest.

Equivalence (alternate forms) is the correlation between two administrations of parallel forms of the same test. This is the best index of test reliability.

Internal consistency is calculated using the Spearman-Brown formula based on a split-half techniques that compares two equivalent halves of a test using odd vs. even numbered items. Another method involves Cronbach's alpha that compares the variance of each item to total test variance.

Rational equivalence uses the Kuder-Richardson (KR) 20 provides relatively conservative estimates of the coefficient of equivalence. KR formula 21 is less accurate, but simple to compute.

Decision-consistency, which is the average of squared deviation from the established mastery level, is used with criterion-referenced tests.

The most popular approach for calculating the reliability of criterion-referenced tests involves the consistency of mastery-non-mastery decision making over repeated measures with one test or randomly parallel tests. A parallel test has more rigid requirements that involves the matching of item pairs and total test score.

Swaminathan, Hambleton, and Alginia (1974) described a decision-consistency procedure to determine reliability based on the proportion of individuals consistently classified as master/master and non-master/non-master on pretest-posttest scores. Table 1 lists test scores for the 10 staff members. The same 10-item test was administered to each person twice (Test 1 and Test 2), and a score of 8 was used to determine mastery.

Table 1
Pretest-Posttest Scores

Trainee	Test 1	Test 2
1	8	10
2	8	8
3	8	8
4	8	9
5	8	8
6	8	7
7	8	6
8	7	8
9	6	7
10	4	5

Table 2 summarizes mastery-nonmastery pretest and posttest outcomes for the scores in Table 1. Trainees #1 through #5 were test 1/test 2 masters, so 5 is entered in cell 1. Trainees #6 and #7 were test 1 masters and test 2 non-masters, so 2 is entered in cell 2. Trainee #8 was a test 1 nonmaster and test 2 master, so 1 is entered in cell 3. Trainees #8 through 10 were non-masters on tests 1 and 2, so 3 is entered in cell 4.

Table 2
Mastery-Nonmastery Scores

	Posttest		
Pretest	Master	Nonmaster	Total
Master	5	2	7
Nonmaster	1	3	4
Total	6	5	11

$p = 5/10 + 3/10 = .80$

The predicted reliability *p is* based on the proportion of staff consistently classified as masters and non-masters is .80. The correlation between a test and the criterion is never higher than the square root of the product of the reliability of the test and the reliability of the criterion variable. All else being equal with tests that measure the same thing, use the test with the highest reliability.

Item-Objective Congruence

The first step in validating a criterion-referenced test is to establish item-objective congruence. It answers the question: How well does the item measure the objective? Match each item to an objective, and assess how well the item measures the performance stated in the objective. The number of items included for an objective should reflect the relative importance of the objective and the training time allocated to the objective. It is not necessary to have a test item for each objective if the total number of items adequately samples the training curriculum.

Table 3 summarizes the criteria that are used to establish item-objective congruence.

Table 3
Item-Objective Congruence

Criterion	Question
Behavior	Is the behavior described in specific, measurable terms?
Content	Does the objective match specific content in the curriculum?
Hierarchical classification	Are the objectives listed in a logical hierarchy?
Level of specificity	Is the level of specificity appropriate for the trainees?
Congruence	Does the item match the objective on the preceding criteria?

Item Revision

Analyzing the pattern of responses for distracters is an effective way to determine the effectiveness of distracters. The following procedures will improve the process of item revision.

> Ideally, trainees should answer all pretest questions incorrectly and all posttest questions correctly.

If a majority of students miss an item, it does not necessarily have to be changed, but check it for accuracy and find out if the material was covered during training.

Revise or eliminate items answered correctly by more than 80 percent of examinees on the pretest or incorrectly less than 30 percent on the posttest.

Rewrite or eliminate posttest items that correlate less than .15 with total test score. These items probably do not measure the same domain of competencies as the entire test or may be confusing.

Eliminate or replace distracters that are not chosen by any examinees.

Prepare tables that summarize gains or losses by item based on trainee pretest/posttest scores. This information may indicate that the pretest is too easy. Consequently, posttest scores may be high because many trainees can answer questions correctly before they begin training. This may indicate that items are poorly written or that the curriculum includes content that trainees learned in other settings.

Discrimination indexes should be positive for correct answers and negative for incorrect answers.

Distracters that are never or infrequently selected should be revised or eliminated (Millman & Greene, 1993).

Report total test reliability first and remove each item from the test and calculate test reliability excluding that item. This procedure generates tests with the highest possible reliability (Thompson & Levitov, 1985, p.167).

Calculate discrimination indexes for each distracter.

Determine the percentage of trainees who answer each item correctly on the pretest and posttest. If too large a percentage answer a pretest item correctly, examine the item, the related objective, and corresponding section of curriculum and make appropriate changes.

More trainees in the uninstructed group should select each distracter than in the instructed group.

At least a few uninstructed students should choose each distracter.

No distracter should be selected by as often by the instructed group as the correct answer.

If more trainees answer an item correctly on the pretest than on the posttest, it is a weak item that should be revised or eliminated, or perhaps it is material that was taught improperly.

Examine the percentage of trainees who select each distracter. Patterns of incorrect responses can reveal misunderstandings, ambiguity, lack of knowledge, guessing or an incorrect response on the answer sheet.

Interview the criterion group to identify problem areas on the test using questions listed in Table 4.

Table 4
Questions for Criterion Group on Sources of Error

Error	Question
Item ambiguity	Did an item seem confusing?
Cueing	Did a question give clues about which answer was correct?
Miskeyed answers	Do you disagree with any of the correct answers?
Inappropriate vocabulary	Were you unfamiliar with any of the words used?
Unclear items	Did an item have two correct answers or no correct answer?
Unclear instructions	Were the instructions clear and easy to understand?

Table 5 compares responses on a single item for the pretest and posttest scores for a group of 25 trainees. The responses indicate that on the pretest comparable numbers of trainees selected each distracter, while on the posttest 22 (88%) trainees selected the correct answer. This pattern indicates that the item was appropriately difficult for untrained people, but that most trainees successfully answered the item after they completed training.

Table 5
Good Pattern of Responses

Option	Pretest	Posttest
A	5	0
B	6	6
C	4	4
*D	7	7
E	3	3

* = correct response

Table 6 illustrates a poor pattern of responses on a question that is probably too easy.

Table 6
Pattern of Responses for an Easy Question

Option	Pretest	Posttest
A	0	0
B	0	0
C	5	0
*D	17	25
E	3	0

* = correct response

Table 7 is a response pattern that indicates a strong possibility of a question with two correct answers (C or D).

Table 7
Misleading Question with Two Possible Answers

Option	Pretest	Posttest
A	6	0
B	3	2
C	6	11
*D	17	12
E	5	0

* = correct response

Table 8 presents a different perspective from item difficulty can be reviewed. A sign of + indicates that a trainee answered a question correctly, while a sign of – indicates an incorrect answer. Many more trained than untrained persons should answer a question correctly. Items that deviate from this pattern should be examined carefully and either revised or omitted.

Table 8
Correct and Incorrect Responses by Item

Instructed						Uninstructed					
Trainee	1	2	3	4	5	Trainee	1	2	3	4	5
1	+	+	+	-	-	1	-	-	-	+	-
2	+	+	+	+	+	2	-	-	-	-	-
3	+	+	+	+	+	3	+	-	-	-	-
4	+	+	+	+	+	4	-	-	-	-	+
5	-	-	-	+	-	5	+	+	+	+	-

As noted above, after the pilot test is completed, review the test and specific items with the people who have completed the test.

Item Difficulty

Item difficulty is the percentage of people who answer an item correctly. It is the relative frequency with which examinees choose the correct response (Thorndike, Cunningham, Thorndike, & Hagen, 1991). It has an index ranging from a low of 0 to a high of +1.00. Higher difficulty indexes indicate easier items. An item answered correctly by 75% of the examinees has an item difficult level of .75. An item answered correctly by 35% of the examinees has an item difficulty level of .35.

Item difficulty is a characteristic of the item and the sample that takes the test. For example, a vocabulary question that asks for synonyms for English nouns will be easy for American graduate students in English literature, but difficult for elementary children. Item difficulty provides a common metric to compare items that measure different domains, such as questions in statistics and sociology making it possible to determine if either item is more difficult for the same group of examinees. Item difficulty has a powerful effect on both the variability of test scores and the precision with which test scores discriminate among groups of examinees (Thorndike, Cunningham, Thorndike, & Hagen, 1991). In discussing procedures to determine minimum and maximum test scores, Thompson and Levitov (1985) said that

> Items tend to improve test reliability when the percentage of students who correctly answer the item is halfway between the percentage expected to correctly answer if pure guessing governed responses and the percentage (100%) who would correctly answer if everyone knew the answer. (pp. 164-165)

Item difficulty is calculated by using the following formula (Crocker & Algina, 1986).

$$\text{Difficulty} = \frac{\text{\# who answered an item correctly}}{\text{Total \# tested}} \quad \text{X } 100$$

For example, assume that 25 people were tested and 23 of them answered the first item correctly.

$$\text{Difficulty} = \frac{23}{25} \text{ X } 100 = .904$$

In norm-referenced tests, the optimal level of difficulty depends on the number of test items and the chance score. The following formula calculates the optimal difficulty level.

$$\text{Optimal difficulty} = \text{Chance Score} + \frac{\text{Perfect score} - \text{chance score}}{\text{Number of options}} \text{ X Number of items}$$

Table 9 lists the optimal difficulty levels for items with different number of options.

Table 9
Optimal Difficulty Levels for Items with Different Options
(for tests with 100 items)

Number of Options	Optimal Difficulty Level
2	.75
3	.67
4	.63
5	.60

A simple way to calculate the ideal difficulty level is to identify the point on the difficulty scale midway between perfect (100 percent) and chance-level difficulty (25 percent for items with four options). The optimal difficulty level is, therefore, 62.5 percent for 4-option items (Thompson & Levitov, 1985). On criterion-referenced tests, however, the optimal difficulty level for items should be very low for instructed groups (less than .30) and very high for uninstructed groups (more than .80).

Item Discrimination

Item discrimination compares the number of high scorers and low scorers who answer an item correctly. It is the extent to which items discriminate among trainees in the high and low groups. The total test and each item should measure the same thing. High performers should be more likely to answer a good item correctly, and low performers more likely to answer incorrectly. Scores range from –1.00 to +1.00 with an ideal score of +1.00. Positive coefficients indicate that high-scoring examinees tended to have higher scores on the item, while a negative coefficient indicates that low-scoring students tended to have lower scores. On items that discriminate well, more high scorers than low scorers will answer those items correctly.

To compute item discrimination, a test is scored, scores are rank-ordered, and 27 percent of the highest and lowest scorers are selected (Kelley, 1939). The number of correct answers in the highest 27 percent is subtracted from the number of correct answers in the lowest 27 percent. This result is divided by the number of people in the larger of the two groups. Th percentage of 27 percent is used because "this value will maximize differences in normal distributions while providing enough cases for analysis" (Wiersma & Jurs, 1990, p. 145). Comparing the upper and lower groups promotes stability by maximizing differences between the two groups. The percentage of individuals included in the highest and lowest groups can vary. Nunnally (1972) suggested 25 percent, while SPSS (1999) uses the highest and lowest one-third.

Wood (1960) stated that

> When more students in the lower group than in the upper group select the right answer to an item, the item actually has negative validity. Assuming that the criterion itself has validity, the item is not only useless but is actually serving to decrease the validity of the test. (p. 87)

The higher the discrimination index, the better the item because high values indicate that the item discriminates in favor of the upper group which should answer more items correctly. If more low scorers answer an item correctly, it will have a negative value and is probably flawed.

A negative discrimination index occurs for items that are too hard or poorly written, which makes it difficult to select the correct answer. On these items poor students may guess correctly, while good students, suspecting that a question is too easy, may answer incorrectly by reading too much into the question. Good items have a discrimination index of .40 and higher; reasonably good items from .30 to .39; marginal items from .20 to .29, and poor items less than .20 (Ebel & Frisbie, 1986).

Discrimination Coefficients

Three correlational methods, including point biserial correlation, biserial correlation and phi coefficient, are used to determine item discrimination. Both methods measure item discrimination by calculating the association between two variables one that is dichotomous and the other continuous.. A dichotomous variable is collapsed into two levels (e.g., high/low; right/wrong; 0/1) and assumes that the collapsed dichotomous variable is continuous (Vogt, 1999). These coefficients have an advantage over the discrimination index because they use every examinee to calculate the coefficient, while only 54% (27% higher + 27% lowest) are used for the discrimination index.

Point-biserial correlation shows how much predictive power an item has and how the item contributes to predictions by estimating the correlation between each test item and the total test score. The statistic is useful for examining the relative performance of different groups or individuals on the same item.

Point-biserial correlation is a product-moment correlation. A moment is a standard score deviation about a mean of zero, and one standard score from the mean as the first deviate. Squared deviates are the second moment. Cubed deviates are the third, and so on. Items with higher point-biserial correlations are more highly discriminating, while those with lower point-biserial correlations are less discriminating. Test developers either drop or revise items with negative point-biserial correlations (Osterlund, 1998).

Biserial correlation is similar to the point-biserial correlation. It shows the extent to which items measure attributes included in the criterion. It estimates the Pearson product-moment correlation between the criterion score and the hypothesized item continuum when the item is dichotomized into right and wrong (Henrysson, 1971). It also describes the relationship between scores on a test item (e.g., "0" or "1") and scores on the total test for all examinees (Ebel & Frisbie, 1986). It differs from point-biserial correlation because it assumes that both variables are inherently continuous, rather than classifying one of the variables as a true dichotomy (Osterlund, 1998).

Henrysson (1971) suggested that point biserial correlation is a better measure of predictive validity than biserial correlation because it favors items of average difficulty and is a combined measure of item-criterion relationship and difficulty.

Phi coefficient yields an estimate between –1.00 and +1.00. It differs from point biserial and biserial correlations by assuming a genuine dichotomy in both correlated variables. It is the degree of association between an item and a criterion (e.g., trained/untrained; demographic characteristic) (Osterlund, 1998).

On criterion-referenced tests reliability is calculated by comparing pretest/posttest scores for the same group of persons or differences between instructed and uninstructed criterion groups. For criterion-referenced tests, the index is calculated in several ways.

> Pretest-posttest difference: proportion answering an item correctly on posttest minus the proportion answering correctly on pretest).
>
> Trained-untrained group difference: proportion in instructed group who answered correctly minus proportion in uninstructed group who answered correctly).

Individual gain: proportion who answered incorrectly on pretest and correctly on posttest).

Net gain: proportion who answered incorrectly on pretest minus proportion who answered incorrectly on both occasions

Table 10 illustrates how item discrimination is calculated.

Table 10
Item Discrimination for Trained and Untrained Groups

Objective	Item	% Correct (trained)	% Correct (untrained)	Discrimination
1.1	1	100%	58%	.42
1.2	2	91%	54%	.37
1.3	3	76%	86%	-.10
1.4	4	100%	95%	.05
1.5	5	82%	8%	.79

For norm-referenced and criterion-referenced test, the criteria are the same. The accepted practice is to omit or revise items with a negative discrimination indexes or indexes ranging from 0 to .30 and to retain items with indexes higher than .35.

Instructional Sensitivity

Cox and Vargas (1966) introduced the concept of instructional sensitivity IS) to measure the effect of training on performance. Although IS includes four theoretical contexts - classical, Bayesian, item response, and criterion-referenced - this discussion focuses on the instructional sensitivity of criterion-referenced tests.

IS is a useful concept in competency-based, criterion-referenced instruction in which mastery is an issue. In systematic instruction, test results would reflect what trainees learned, and trained people should perform better than untrained people. Differences in pretest and posttest means are a measure of instructional effectiveness (Roid & Haldanya, 1982).

The pre-to-post difference index (PPDI) is assessed using the pretest/posttest performance of a single group or the scores of two randomly selected groups of training and untrained individuals. Table 11 summarizes three different examples of this process described by Haladyna (1994).

Table 11
Examples of PPDI for 4-option items

Item	Pre-instruction	Post-instruction	PPDI
A	40%	80%	40%
B	40%	40%	0
C	90%	90%	0
D	12%	92%	80%

On the pretest for 4-option, criterion-referenced items, most trainees should answer incorrectly substantially below the probability expectation of 25 percent. High posttest scores exceeding 80 percent should be an expectation. If PPDI results are unsatisfactory, trainers should review the curriculum, objectives, and test item to determine what must be improved or refined.

> Item A is moderately difficult because 60 percent of the trainees answered it incorrectly. The improvement of 40 percent represents a moderate increase in performance.
>
> Item B suggests ineffective or inadequate instruction, an excessively difficult or confusing item, or an item unrelated to the curriculum.
>
> Item C illustrates an item that is too easy either because trainees already know the material or that the item was poorly written and cued the correct answer.
>
> Item D represents an item that tested unfamiliar content that most people mastered during training.

Standard Setting

A standard is a performance score that a person must achieve to be classified as master/non-master, pass/fail, or a proficiency level such as high, average, and low. Performance scores are typically based on the judgment of experts who specify the criteria that will be used to sort examinees into categories. If testing is used to determine whether trainees will be retained or dismissed, establishing fair, well-defined performance standards improves hiring practices. Procedures used to establish standards should be publicized as a matter of record. Setting standards for employment tests is critical for several reasons:

Legal - avoid challenges regarding employment decisions.

Ethical - follow moral obligation to employ best candidates.

Pragmatic - employ competent people to meet organizational goals.

Where a standard is set depends on the purpose of the test. Norm-referenced tests are designed to yield test scores that approximate a normal curve distribution. The *Wechsler Intelligence Scale for Children* (WISC), for example, has a mean of 100 and a standard deviation of 15, a score of 115 falls one standard deviation above the mean at the 84th percentile.

Mastery tests establish cutoff scores at a level that depends on the purpose of the test. Highly selective programs that admit few applicants, such as medical school or professional basketball, have cutoff scores that eliminate most candidates. Other programs, particularly those that are competency-based, are designed to have most people achieve mastery.

Hambleton (1980) described several methods for establishing standards including Nedelsky (1954), Ebel & Frisbie (1986), and Angoff (1971). Each method uses the judgment of experts.

Nedelsky (1954) asks judges to identify distracters in multiple-choice items that a minimally competent person would eliminate as incorrect. The minimum passing level for an item is the “chance score” which is the reciprocal of the remaining alternatives. Ratings of individual judges are averaged to obtain a standard for the test.

Ebel & Frisbie (1986) has judges classify items within a 3 X 4 grid including four levels of relevance (essential, important, acceptable, and questionable) and three levels of difficulty (easy, medium, and hard). Then, judges estimate what percentage of items in a cell can be answered by minimally qualified examinees. This percentage is multiplied by the number of test items in each cell, and the standard is obtained by dividing the sum of all cells by the total number of test items.

Angoff (1971) selects a representative panel from a population of qualified judges. Panel members develop, or are given, a definition of borderline competence, which is a clear conceptualization of a minimally competent candidate. Judges review this definition and discuss what constitutes borderline or minimal knowledge and skills.

Judges consider each item on the test and estimate the probability that a minimally competent trainee would answer the item correctly. Probabilities for each judge are summed which is the judge's estimate of the total score a minimally competent trainee would achieve. The final step is to average the sums of probabilities for all judges which provides an estimate of the score that should be earned by minimally competent trainees.

Generalizability Theory

Generalizability theory replaces a true score with a "universe" score. A universe score is defined as a sample from a universe of scores that would exist if measurements were obtained under all admissible conditions. It may include facets constituted by random or fixed modes of sampling, but it must include random sampling of measurement conditions (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

For example, if people were crossed with a random sample of test items, the facet "items" could be nested within the levels of a fixed facet such as instructional objectives, or crossed with levels of a random facet such as settings. After data are collected, ANOVA determines the mean square for each source of variation. Numerical values for observed mean squares are successively inserted into the equations for the expected mean squares providing estimates for variance components corresponding to all sources of variation (Allal, 1990).

The generalizability model yields three major parameters after variance components are estimated.

> Universe-score variance is the variance of expected scores in the universe of admissible observations. It reflects systematic variations due to differences among measurement objects.
>
> Relative error variance is sum of weighted variance component estimates corresponding to sources of relative error. Each component is weighted inversely to the number of times its effect is sampled when calculating the average score of one measurement object.
>
> Absolute error variance includes components of relative error, plus components of specific facets formed by random sampling of measurement conditions (Allal, 1990).

Latent Trait / Item Response Theory

Latent trait theory is also called item response theory (IRT). IRT measures underlying abilities called “latent traits” which are theoretical attributes (or abilities) that define test performance independently and completely. IRT assumes the existence of an underlying scale with equal-interval units that measure a single (or unidimensional) latent trait. Height is a unidimensional latent trait because the concept cannot be seen, but it can be measured. A psychological attribute is less likely to be unidimensional (e.g., intelligence includes multiple factors such as reading ability, vocabulary, and motivation).

One-parameter IRT models are called Rasch models after the Danish mathematician George Rasch (1960). Rasch scaling, which is used with IRT, measures the extent to which items contribute to unidimensionality. Assume that 10 people complete a 5-item test to measure motivation level. A Rasch scale ranks raw scores from lowest to highest for the 10 people and ranks the 5 items from least to most difficult to compare the performance of each ability according to each level of item difficulty. This process makes it possible to tailor tests for individual abilities by using items appropriate to each person's ability level. Most large test publishers, state departments of education, and industrial and professional organizations use IRT construct and equate tests (Hambleton & Murray, 1983).

A latent trait is a construct that represents a single, underlying ability determined by a person's test performance (Lord & Novick, 1968). It is an attribute that accounts for the consistency of person's responses (Wainer & Messick, 1983).

> The probability that a person will answer an item correctly is assumed to be the product of an ability parameter pertaining only to the person and a difficulty parameter pertaining only to the item. (Loevinger, 1965, p. 151)

A latent trait is visualized as a monotonically increasing curve on which persons and test items can be placed according to their ability and the level of difficulty of items. Individual responses to specific test items determine position on the continuum, which is a numerical scale (Hulin, Drasgow & Parsons, 1983).

Alderson, Clapham, and Wall (1995) described three IRT models:

> One-parameter models place item difficulty and person's ability level on the same continuum representing different manifestations of the same parameter with item difficulty directly comparable to ability. The one-parameter model is relatively simple and provides reliable estimates with as few as a 100 subjects.
>
> Two-parameter models yield higher accuracy because they include differences in item discriminations, but its increased complexity requires at least 200 subjects.

> Three-parameter models add guessing and other deviant behaviors to the analysis, but require samples of 1,000 subjects.

IRT assumes that each item has a fixed difficulty level. A person's ability level is based on the probability that the person responds correctly to an item or set of items. It is similar to cluster analysis and uses multivariate categorical data to classify cases into categories or latent traits based on symptoms, attitudes, or behavior (Day, 1969; Wolfe, 1970).

IRT can be used with binary, ordered-category, and Likert-scale data, or nominal, but not ordinal data. The results can classify people into categories such as master and non-master or expert, intermediate, and novice.

Differential Item Functioning (DIF)

Test bias is a serious concern, particularly if test results determine whether a person will be hired, retained, or promoted. For example, critics have argued that some verbal reasoning questions on the *Scholastic Aptitude Test* (e.g., the analogy "runner is to marathon as oarsman is to regatta") are biased in favor of privileged students and against those from a disadvantaged background (Camilli & Shepard, 1994). To compensate for this type of bias, the Educational Testing Service (1999) uses differential item functioning (DIF).

DIF identifies test questions that give one groups of test takers an unfair advantage over another. It assumes that test takers who have approximately the same level of knowledge, as measured by total test scores, should perform in a similar manner on individual test questions regardless of sex, race or ethnicity. Relevant focal groups (e.g., minority, female) are identified. Trainees are matched by test scores and compared on how well they performed on individual test items. ETS reviews questions on which focal groups perform differently are reviewed for possible bias.

DIF scores are presented as differences on a scale that ranks test questions according to difficulty. A negative value means that the test question is more difficult for the focal group, while a positive value indicates that it is more difficult for the reference group (males or whites). The higher the number, the greater the difference between matched groups.

Questions that are harder for focal groups are identified. A DIF panel of experienced professionals decides whether flagged questions relate to what the test was designed should measure or is unfair to members of the focal group. Unfair questions are modified or eliminated.

DIF is an important consideration for organizations that use tests to select staff, such as departments of social services and education. Two landmark Supreme Court decisions including *Griggs v. Duke Power* 401 U.S. 424 (1971) and *Hazelwood School District v. United States* 433 U.S. 299 (1977) held that a test is biased if it predicts differences between protected groups when no real difference can be proven. Subsequent rulings held that individual test items must also withstand tests of item bias. These decisions established an 80 / 20 guideline which maintains that for a test or item to be considered unbiased, the performance of protected groups must fall within 80 percent of the majority group (Yegidis & Morton, 1999). Obviously, this issue is a serious concern for training organizations.

Conclusion

This monograph described major concepts related to item analysis including validity, reliability, item difficulty, and item discrimination, particularly in relation to criterion-referenced tests. The paper discussed how these concepts can be used to revise and improve items and listed suggestions regarding general guidelines for test development. The paper concluded with a brief discussion of standard setting, latent trait theory, item response theory, and differential item functioning.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Allal, L. (1990). Generalizability theory. In H. J. Walberg & G. D. Haertel (Eds.). *The international encyclopedia of educational evaluation*. New York: Pergamon Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement and Education. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*, 201-238.

American Psychological Association, (1966*). Standards for educational and psychological tests and manuals*. Washington, D.C.: Author.

Anastasi, A. (1954). *Psychological testing*. New York: Macmillan.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Macmillan.

Angoff, W. W. (1971). Scales, norms, and equivalent scores. In R. L Thorndike (Ed.). *Educational Measurement* 2nd Ed. Washington, D.C.: American Council on Education.

Berk, R. A. (1980). Item analysis. In R. A. Berk (Ed.). *Criterion-referenced measurement: The state of the art*. Baltimore: The Johns Hopkins University Press.

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Vol. 4. Thousand Oaks, CA: Sage Publications.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity in the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cox, R. C., & Vargas, J. (1966). *A comparison of item selection techniques for norm-referenced and criterion-referenced tests.* Pittsburgh: University of Pittsburgh Learning Research and Development Center.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1949*). Essentials of psychological testing*. New York: Harper & Row.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.). *Intelligence: measurement theory and public policy* (pp. 147-171). Urbana: University of Illinois Press.

Cronbach, L. J., Gleser, G., Nanda, H., & Rajaranam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1954). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Cureton, E. F. (1951). Validity. In E. F. Lindquist (Ed.). *Educational measurement* (1st Ed.). Washington, DC: American Council on Education.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika, 56*, 463-474.

Educational Testing Service. (1999, August 10). What's the DIF? Helping to ensure test question fairness. *research@ets.org*. Princeton, NJ: The Educational Testing Service.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th Ed.). New York: MacMillan.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-439.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.). *Educational measurement*. (3rd Ed.). New York: Macmillan Publishing.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.). *Criterion-referenced measurement: The state of the art*. Baltimore: The Johns Hopkins University Press.

Hambleton, R. K., & Murray, L. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.). *Applications of item response theory*. Vancouver, British Columbia: Educational Research Institute of British Columbia.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R.L. Thorndike (Ed.), *Educational Measurement,* Washington, DC: American Council on Education.

Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983). *Item response theory*. Homewood, IL: Dow-Jones Irwin.

Kelley, T. I. (1939). The selection of upper and lower gru0ps for the validation of test items, *Journal of Educational psychology, 30,*(1), 17-24.

Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd ed.). New York: Holt, Rinehart, and Winston.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Reviews, 72*, 143-55.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955-966.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In H. Wainer, & H. I. Braun (Eds.). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational measurement* (3rd Ed.). New York: American Council on Education and Macmillan.

Millman, J., & Greene, J. (1993). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed.). New York: American Council on Education and Macmillan.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Nunnally, J. C. (1972). *Educational measurement and evaluation* (2nd Ed.). New York: McGraw-Hill.

Osterlund, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2nd Ed.). Boston: Bluwer Academic Publishers.

Popham, W. J. (1981). *Modern educational measurement*. Englewood Cliff, NJ: Prentice-Hall.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing.* New York: Harcourt Brace Jovanovich.

Sax, G. (1989). *Principles of educational and psychological measurement and evaluation* (3rd Ed.). Belmont, CA: Wadsworth.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology. 15,* 201-93.

SPSS. (1999). Item analysis. *spss.com*. Chicago: Statistical Package for the Social Sciences.

Subkoviak, M. J. (1980). Decision-consistency approaches. In R. A. Berk (Ed.). *Criterion-referenced measurement: The state of the art*. Baltimore: The Johns Hopkins University Press.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11,* 263-267.

Thompson, B., & Levitov, J. E. (1985). Using microcomputers to score and evaluate test items. *Collegiate Microcomputer, 3*, 163-168.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education* (5th Ed.). New York: MacMillan.

Vogt, W. P. (1999). *Dictionary of statistics and methodology: A non-technical guide for the social sciences* (2nd Ed.). Thousand Oaks, CA: Sage Publications.

Wiersma, W. & Jurs, S. G. (1990). *Educational measurement and testing* (2nd Ed.). Boston, MA: Allyn and Bacon.

Wainer, H. & Messick, S. (1983). *Principles of modern psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research, 5*, 329-350.

Wood, D. A. (1960). *Test construction: Development and interpretation of achievement tests*. Columbus, OH: Charles E. Merrill Books, Inc.

Yegidis, B., & Morton, T. D. (1999, March). Item bias and CPS assessments. *Ideas in Action*. Atlanta: Child Welfare Institute.

BSC
Buffalo State College
State University of New York

CDHS
Center for Development
of Human Services

Center for Development of Human Services
Research Foundation of SUNY
State University College at Buffalo
1695 Elmwood Avenue
Buffalo, New York 14207-2407
716.876.7600 (Voice)
716.876.2201 (Fax)