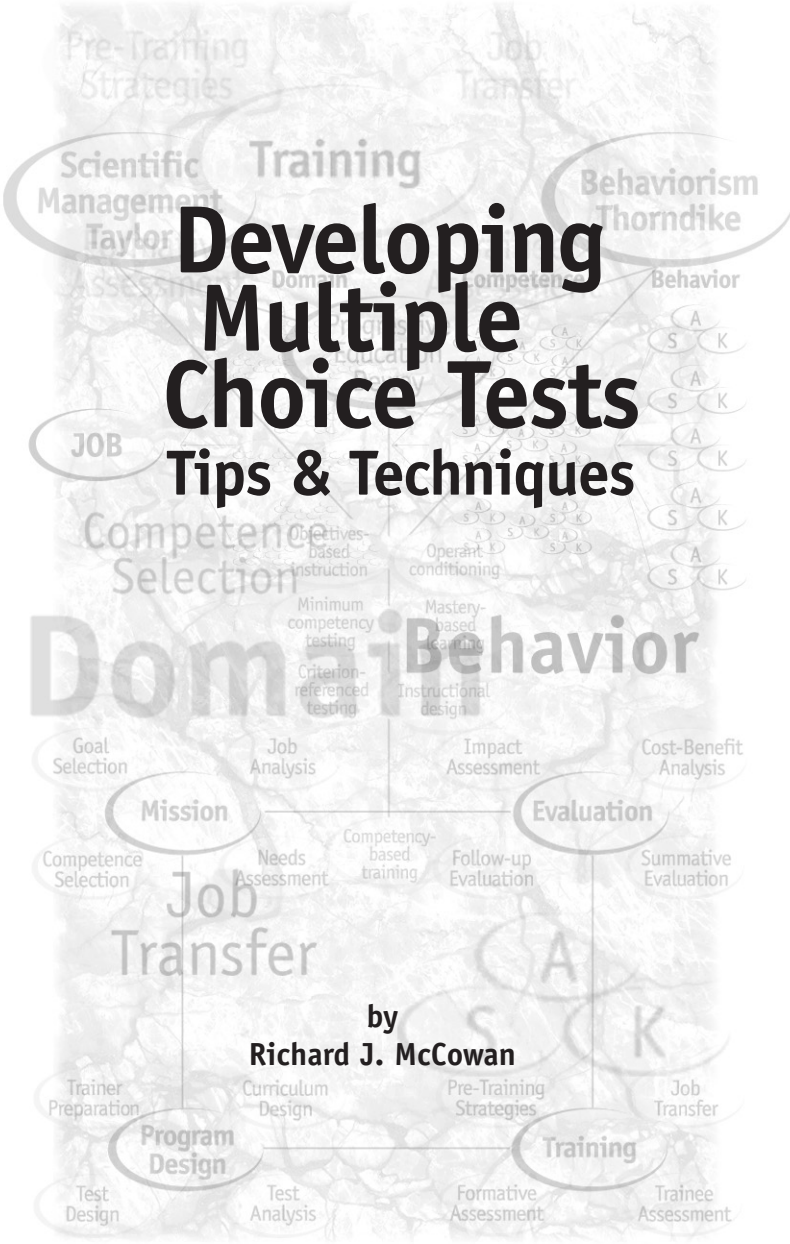
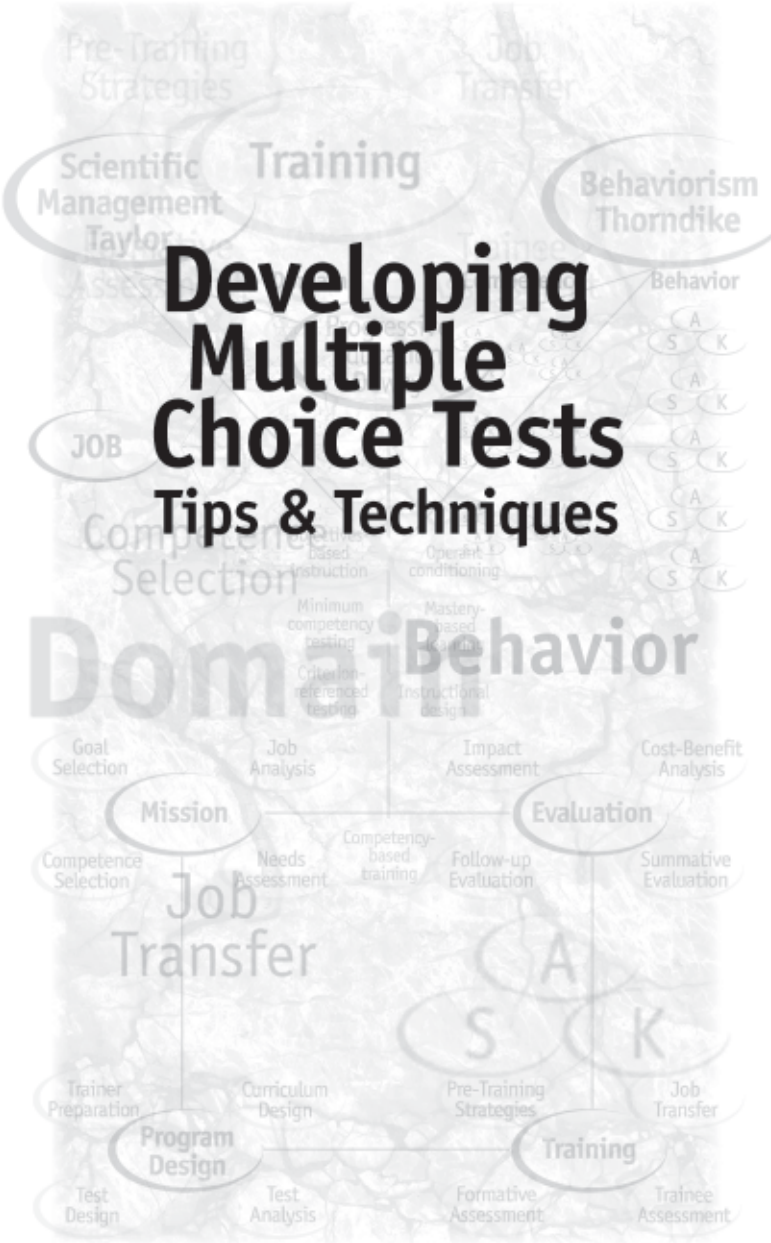


Developing Multiple Choice Tests

Tips & Techniques

by
Richard J. McCowan



Developing Multiple Choice Tests

Tips & Techniques

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

Social work literature provides limited advice for test developers. A review of professional social work journals covering the last 20 years revealed no published articles on using or developing multiple choice examinations. Since neither undergraduate or graduate social work programs emphasize psychometrics and item-writing, it is not surprising that many social services trainers lack skill in item-writing.

Trainers have legitimate reasons for questioning multiple choice tests. Professional journals and popular media publish horrible examples of poorly constructed, confusing questions. Trainers complain that emphasizing use of multiple choice items encourages trainees to memorize trivial details to the detriment of more important aspects of the curriculum. They note that this type of test hinders the development of creativity and writing skills. They are reluctant to use such impartial, mechanistic measurement tools to test the performance of future social workers and dismiss these items as "multiple guess tests."

Proponents of multiple choice tests disagree. They cite the efficiency and reliability of the tests, particularly if computer scoring is available. They note that properly constructed multiple choice tests assess a full range of cognitive and attitudinal domains. Multiple choice tests are also a clear choice when it is necessary to assess the performance of large numbers of trainees.

This paper will not resolve the debate because supporters of both viewpoints raise cogent arguments. Multiple choice items provide an efficient, valid way to test most dimensions of training programs. In New York State, elementary children are tested to determine if they are minimally competent in basic skills, and high school students must pass basic skills tests to graduate from high school and to be admitted to college and graduate school. Obtaining a professional license, including social work, is determined to a major extent by performance on multiple choice examinations. Multiple-choice items are effective in testing most aspects of competency-based training programs.

Most of the suggestions in this monograph are supported by psychometric research, while other recommendations are based on good judgment. Consistent use of these suggestions will increase the validity and reliability of multiple-choice tests, and data from these tests can be used to improve training programs.

Test Instructions

Write complete test instructions that are brief and clear.

Illustration

> Use the scan sheet for all responses to this test, except where indicated.
>
> Print your Social Security number in the spaces provided for Identification Number and darken the spaces for each number using a #2 pencil.
>
> Then, print the Special Code in the spaces provided and darken the space for each number. This Special Code is printed on the right hand corner of this page.
>
> After completing this information, please answer all questions on this test by darkening the appropriate space on the scan sheet. Darken only one space per question. Erase changed answers completely and do not make any stray marks on the sheet.

Read test instructions to the trainees and answer questions regarding the test, but, preferably, write instructions that make it unnecessary for them to ask any questions.

Use identification numbers (preferably Social Security numbers) to identify individual trainees.

Use a special code to identify the training session during which the test was used.

Limit demographic information to items related to the purpose of the test.

Avoid irrelevant, intrusive demographic items (e.g., gender, marital status, and age).

Use computer scan sheets to record trainee responses for efficient, accurate scoring.

Item Stems

State questions precisely and carefully (or, say what you mean).

Poor Example

In recent times how has social casework been more effective than community organization?

This example uses vague terms including "recent times," "social casework," "community organization," and "effective." Rewrite the item using operational terms such as "since 1995" and "managed care" instead of "recent times" and "social casework."

Include the main concept in the item stem with all necessary information to avoid repetitive, excessively long distracters.

Eliminate repetitive, irrelevant material (e.g., "of the following"; "for the most part").

Poor Example

(Question is based on map of the eastern United States describing population changes from 1980-90).

Which of the following best describes why people moved from areas with the greatest out-migration, as indicated on the map?

a. The migrants moved to states that have more desirable, warmer weather most of the year.

b. The migrants moved in greater numbers to the southern and western states.

c. The migrants sought new forms of entertainment, recreation, and employment by moving to states that provided greater opportunities.

d. The migrants moved to states that had less expensive housing.

Good Example

Why did people move from states with the greatest out-migration?

a. warmer weather

b. geographic location

c. recreational and employment opportunities

d. less expensive housing *

Follow standard rules on punctuation and grammar. Unusual or incorrect punctuation and grammar are confusing and provide a poor model of good writing. The following stem is not an interrogative sentence, despite the use of a question mark.

Poor Example

Neglectful parents are likely to learn parenting skills best by doing which of the following?

Good Example

How can neglectful parents best learn effective parenting techniques?

Avoid long item stems and keep the reading level and vocabulary as simple as possible, unless reading ability is being tested. Reading long item stems penalizes trainees who know the content, but read slowly.

Eliminate superfluous information in the item stem.

Poor Example

Which therapist, who was thought by some people to have an erratic, bohemian life style, was the major proponent of Gestalt therapy?

a. Carl Rogers

b. Albert Ellis

c. Robert Carkhoff

d. Fritz Perls *

Good Example

Which therapist was the major proponent of Gestalt therapy? (same distracters)

Don't test knowledge of trivial information.

Poor Example

On which day was Sigmund Freud born?

a. March 21, 1856

b. April 6, 1856

c. September 9, 1956

d. October 10, 1956

If dates related to a theory are important, use broad periods of time that relate the theory to a broad, historical framework.

Good Example

When was Freud was most productive?

a. 1800-1850

b. 1885-1920 *

c. 1925-1950

d. 1950-1975

Avoid using questions based on lists of procedures that vary depending on the authority used.

Poor Example

What is the third step in problem-solving?

a. develop a range of possible alternatives

b. obtain useful information about alternatives

c. brainstorm alternative strategies

d. develop feelings of self-efficacy

Criticism: Each distracter was taken from a published problem-solving model, so each is correct. It is more important that trainees learn the general way in which problems are solved, rather than the specific steps in a unique model developed for a training session.

Keep the reading level and vocabulary as simple as possible, unless reading ability is being tested. Tests with long reading passages penalize trainees who know the content, but who read slowly or poorly.

Distracters

Randomly vary the placement of the correct option.

Format distracters vertically, as demonstrated by sample items in this monograph. Although horizontally formatted distracters take less space, a vertical format is easier to read.

Poor Example

Which item format is best to measure cognitive knowledge?

a. true-false b. essay c. multiple-choice* d. matching

Good Example

Which item format is best to measure cognitive knowledge?

a. true-false

b. essay

c. multiple-choice *

d. matching

Use clear, well-defined terms and concepts for distracters.

Poor Example

Which theory maintains that alcohol abuse is a behavior learned from parents who use alcohol to deal with their personal problems?

a. holistic

b. psychological

c. sociological

d. physicalCriticism:

Criticism: The item uses four vague distracters, each of which might be correct, depending upon how each term was defined during training.

Use three or four good distracters for each item to limit the effects of guessing. Research suggests that a small advantage re-

sults by increasing the number of options, but adding a poor distracter weakens an item.

Vary the number of distracters, if necessary. Three or four options are usually desirable, but some questions may have only two possible answers.

Good Example

What will probably result to the level of a child's negative behavior if it is ignored?

a. increase

b. decrease

c. remain the same

Eliminate non-performing distracters, such as those seldom selected, by examining patterns of response. Every distracter should be a plausible answer.

Increase item difficulty by making distracters more homogeneous. Use terms such as "most," "best," "primary," or "least" in the item stem if more than one answer is correct.

Don't use silly or implausible distracters. One poor distracter increases the chance that a trainee can guess the answer.

Poor Example

Who was a major figure in the settlement house movement in the US?

a. Jane Addams

b. Bertha Reynolds

c. Mary Richmond

d. Babe Ruth

Good Example

Who was the major figure in the settlement house movement in the US?

a. Jane Addams *

b. Bertha Reynolds

c. Mary Richmond

d. Helen Harris Perlman

Avoid overlapping alternatives.

Poor Example

Based on the concept of danger and threat to others or self, a court may find people mentally ill when they endanger . . .

a. others

b. themselves

c. themselves or others

d. themselves but not others

Good Example

Based on the concept of danger and threat to others or self, when may a court find people mentally ill?

a. only when they endanger others

b. only when they endanger themselves

c. when they endanger themselves or others *

Don't use "all of the above" or "none of the above" as

distracters. Trainers often use these phrases as correct answers when they cannot think of an incorrect one. If two distracters are correct, "all of the above" is obviously the proper choice.

Poor Example

What behavior therapy technique can help people overcome fear of the opposite sex?

a. systematic desensitization

b. assertion trained

c. modeling

d. behavior rehearsal

e. all of the above

Good Example

What behavior therapy technique establishes a hierarchy of anxiety provoking situations as the first step in counseling?

a. systematic desensitization *

b. assertiveness training

c. modeling

d. behavior rehearsal

Arrange distracters in a logical or numerical order.

Poor Example

During which developmental period is managing frustration a primary task?

a. 7 to 12 years

b. 15 months to 3 years

c. 3 to 6 years *

d. birth to 15 months

Good Example

During which developmental period is managing frustration a primary task?

a. birth to 15 months

b. 15 months to 3 years

c. 3 to 6 years *

d. 7 to 12 years

Don't use complex multiple choice questions in which several distracters are combined into a single distracter.

Poor Example

Effective 10/1/89, in New York State, the rules regarding transfer of resource to qualify for assistance state that:

I. A homestead can be transferred to specific relatives.

II. An A/R must be allowed a reasonable time frame- at least 20 days - to present evidence rebutting presumption of transfer.

III. A client is liable for transfers made by a client's designated power of attorney.

IV. A penalty period is calculated to ascertain the period of restricted coverage.

a. I and II

b. I, II, and III

c. I, I, III, and IV *

d. I, II, and IV

Use test material and terms that are well-known and generally

acceptable in the discipline, rather than items that are textbook or article specific.

Poor Example

What does Kolevzon say about the phi coefficient as a method for testing the relationship between two dichotomous categorical variables?

a. involves assumptions difficult to meet in ordinary testing

b. produces coefficients that exceed unity

c. it is the appropriate method

d. improper for nominal measurement

Good Example

Which statistical technique tests the relationship between two dichotomous nominal level variables?

a. phi coefficient

b. Spearman's *rho* *

c. t-test

d. ANOVA

Grammatical Clues

Write clearly because sloppy, careless writing confuses test takers.

Follow standard rules on punctuation and grammar. Unusual or incorrect punctuation and grammar are confusing and a poor model of clear writing. The following stem is not an interrogative sentence, despite the use of a question mark.

Poor Example

Neglectful parents are likely to learn parenting skills best by doing which of the following?

Good Example

How can neglectful parents best learn effective parenting techniques?

Schedule time for editing and revision, including review by experts.

Write clear, parsimonious items with a single, correct distracter.

Underline or italicize selected words (e.g., "most," "best," "least") for clarity.

Don't give grammatical clues in the test item such as singular or plural verbs that correspond to only one distracter.

Use a question for the item stem to limit grammatical clues.

Poor Example

Among the causes of parental child abuse are

a. low level of parental education

b. pornography

c. type of employment

d. emotional disturbance and substance abuse

Good Example

What contributes most to child abuse?
(same distracters - "d" is correct)

Don't use content clues that give information about the answer.

Poor Example

What power is possessed by a court with appellate jurisdiction?

a. must have a jury

b. power to review and decide appeals *

c. can conduct the original trial

d. can declare laws unconstitutional

Criticism: Remove the words "power" and "appeals" and change the correct choice to "can review decisions of other courts" (Tuckman, 1975, 94-95).

Don't use "always," "never," "all," or "none," particularly in true-false tests, because distracters using these terms are usually false.

Poor Example

What is the major theme in A. Maslow's 1970 book *Motivation and Personality*?

a. meeting physical needs is *always* the major purpose of human striving

b. social factors are *never* important to consider in meeting human needs.

c. basic human needs *never* include rebellion

d. common human needs include social and biological factors

Criticism: Remove specific determiners, such as "always," "all," and "never," from items to make all distracters plausible.

Eliminate information in one item that can be used to answer another.

Poor Example

In script analysis, if miniscripting is a second-by-second process, then

a. game analysis is _____

b. racket analysis is _____

Good Example

In script analysis, what is game analysis?

Don't use sexist language. It is rarely necessary to use masculine or feminine pronouns. As a last resort, use the plural form. If no other recourse is possible, randomly vary the use of gender-related pronouns.

Don't use items that require trainees to insert words deleted from direct quotations from a textbook. This is a misuse of the cloze technique which is used to test reading comprehension. It encourages memorization and is confusing because several correct answers might be inserted in the space.

Avoid items that cannot be completed without information obtained by correctly answering previous items. This penalizes trainees too severely if they do not know the first question.

Poor Example

If $Y + 12 = X - 32$, then

1. $X =$

2. $X^2 =$

Avoid negatively stated item stems. These are confusing and increase the chance of using a double negative. In the rare case when negatively stated items cannot be avoided, underline or italicize the negative term.

Poor Example

Which is not a characteristic of gifted children?

a. emotional stability

b. not awkward

c. not as old as their classmates

d. friendly

Good Example

What is a characteristic of gifted children?

a. emotional instability

b. awkwardness

c. younger than classmates *

d. unfriendly

Avoid using more detail in the stem or distracters than is required to test the concept. The "poor" question listed below would test computational ability, but would confound the process of determining if a trainee understood the concept of a mean score.

Poor Example

What is the mean of these numbers?
212.3479 32/45 1.798643 2,345.67

a. 639.96525

b. 639.96529

c. 639.96533

d. 639.96537

Good Example

What is the mean of these numbers?
2 2 3 4 9

a. 2

b. 3

c. 4 *

d. 5

Include only one defensible option for items that have a single, best answer.

Poor Example

Which is the most serious health problem in the United States?

a. AIDS

b. cancer

c. heart disease

d. tuberculosis

Criticism: Include more information in the stem. For example, "AIDS" would be the correct answer for the question "Which health problem has no effective cure at the present time?", while "heart disease" is the best choice for the question "Which disease most frequently afflicts middle-aged men?"

Make all distracters approximately the same length. Usually the

longest distracter that contains more detail is the correct answer. Include necessary details in the item stem.

Poor Example

Social security income . . .

a. provides an adequate pension

b. reduces savings

c. creates dependency

d. reduces poverty levels if there are other income sources *

Good Example

If a person has other sources of retirement income, what is the effect social security income?

a. provides a better pension *

b. reduces savings

c. creates dependency

d. reduces the poverty level

Don't ask for an opinion if several distracters are correct.

Poor Example

What is the most popular field of practice for social workers?

a. health

b. aging

c. public welfare *

d. child welfare

Good Example

Which field of practice employs the greatest number of social workers?

(same distracters - “c” is correct)

Don’t repeat wording from the stem in the correct response because it provides a clue to the answer.

Avoid using humorous questions. Although humor will probably not have a major effect on test scores, it may distract trainees.

Guessing

Guessing is a serious concern for test developers. Educated persons who read well often "figure out" poorly written items, even if they do not know the material. This tends to increase pretest scores, thus decreasing pretest-posttest differences. Standardized tests, such as the *Scholastic Aptitude Test* and *Graduate Record Examination*, discourage guessing by using a "correction for guessing" formula that subtracts a percentage of incorrect answers from the number of correct answers. Omitted items are not counted as wrong. However, few trainer-developed or criterion-referenced tests correct for guessing.

It is difficult to write good test items. The Educational Testing Service estimates that skilled item writers produce an average of only seven items a day. Since trainers have many responsibilities and limited experience in writing items, they often develop "less than perfect" test items on which the answer can readily be guessed.

Use "unfamiliar with material" as a pretest distracter to reduce guessing and provide a more precise measure of trainee pretest knowledge.

Weighting Scores

Weighting items does not adversely affect test reliability or validity.

Consider assigning two or three points to important items instead of scoring all items equally as one point. Errors on these items would carry a proportional penalty.

Construct tests that reflect the amount of time spent on certain topics and the importance of material covered.

Use anecdotal items with a single best answer to test process skills and higher order thinking.

Illustration

INSTRUCTIONS: Read the excerpt and select the most empathic counselor response. An empathic response is one in which the counselor identifies underlying feelings and reflects content that complements the affective level of the client.

Client: Mary R., 27 years old, high school graduate, unemployed factory worker, fired seven months ago, is applying for Public Assistance. She is single with one child. This is the first counseling session.

"Don't you think that corporations are pretty heartless, even cruel? That is, they hire a person for their own purposes. They don't give a damn about you. Don't you think so?"

Counselor Responses:

a. You're interested in my attitude towards corporations.

b. You feel rather strongly that corporations are all for themselves and the individual who works for them just doesn't count.

c. In other words, you feel that corporations rather consistently violate the integrity of the people they employ.

d. You're depressed, maybe angry, because corporations don't have the proper concern and respect for their employees.

e. You feel that corporations have no hearts, and they hire people only for their own purposes without any concern about the people themselves.

Use anecdotal items with weighted scoring to test higher level skills of evaluation and synthesis. The following example describes different levels of empathy, rated using a seven point scale. This is an efficient method of testing that requires less testing time because each response is a separate item. Consequently, one anecdote represents five test questions.

Illustration

INSTRUCTIONS: Read each excerpt and rate the level of empathy for each counselor response using the seven point rating scale.

1	2	3	4	5	6	7
No attention to surface feelings	Superficial awareness of surface feelings	Minimal recognition of surface feelings	Identifies surface feeling and emotions	Accurately reflects surface feelings	Reflects underlying emotions	Enhances feelings and emotions

Client: Nancy C., 34, married, housewife, three children boys, 5 and 7, daughter, 9. This is the second counseling session.

"My children are getting out of hand. They don't listen to me or my husband unless we threaten them. Who wants to threaten their children all the time? The oldest boy, Jimmy, was really well behaved until last year and then suddenly he's a different kid. He's wild now, always yelling and screaming. Last week I caught him twisting his brother's arm. He really wanted to hurt him!

"I don't know where he gets it from. We're not a violent family. I could see it if I hit the kids all the time. But I don't. Oh, a swat now and then, but I never hurt them. My husband spanks them hard sometimes, but he's not a mean person."

Counselor Responses:

a. "Raising children is a difficult job. It takes a lot of lot of patience and understanding, but corporal punishment, from my point of view, is always wrong."

b. "You're concerned and puzzled by your children's behavior, particularly Jimmy's."

c. "You don't know what to do about their behavior."

d. "Do you know of anything going on with the kids that has occurred recently? Have you talked to them about this?"

Use anecdotal items as indirect measures of values and attitudes.

Rather than asking trainees to assess their own behavior, ask them to evaluate the behaviors of others. Ask trainees to select the best answer or to rate each distracter as demonstrated in the preceding illustration.

Illustration

Client: Jim works in an agency with a photocopy machine where the policy prohibits employees from using the machine for personal business. Twice in the past month, after his supervisor observed him copying magazine articles unrelated to his job, he was warned that this was wrong and that he should not do it again.

What should Jim's supervisor do if he continues to copy personal material?

a. Fire him because he ignored agency policy.

b. Warn him that he will be terminated if he continues to misuse photocopier.

c. Charge him for the copies that he made.

d. Send him a letter describing what will happen if he continues this behavior.

e. Discuss the problem with your supervisor.

f. Ignore it because otherwise he is a good employee.

Instructional Taxonomies

Use items that appropriately cover the full range of knowledge (cognitive), attitudes (affective), and skills (psychomotor) in which people are trained.

Taxonomies, which were originally used in biology to classify life forms, are classifications or sequenced, hierarchical lists that conform to specific rules and follow explicit principles.

In 1956, Benjamin S. Bloom and his associates developed the *Taxonomy of Educational Objectives: Handbook 1: The Cognitive Domain* (know, comprehend, apply, analyze, synthesize and evaluate). At first, the book attracted meager attention, and sales were modest. During the 1960s, as interest in specific educational objectives increased, sales were substantial. The cognitive taxonomy is the most explicit and easiest to measure.

In 1964, David Krathwohl and his co-workers developed a taxonomy for the affective domain (receive, respond, value, organize, and characterize). The affective domain is less precise and more difficult to measure than the cognitive and psychomotor domains.

Other scholars developed psychomotor taxonomies such as motor, perception, physical, nonverbal, and communication (Harrow, 1970; Kibler, Barker, & Miles, 1979).

Sequence cognitive and affective objectives in hierarchical order to maximize learning, but this is not required in the psychomotor domain.

Figure 1 contains the cognitive, affective, and psychomotor taxonomies.

Figure 1

Instructional Taxonomies

Cognitive

Knowledge: recall specific, universal information including methods, processes, abstractions, patterns, structure, and setting (e.g., list, label, state, match, identify).

Comprehension: understand what is communicated by translation, interpretation, and extrapolation (e.g., estimate, infer, predict, translate, illustrate).

Application: use abstractions, particularly in concrete situations (e.g., create, produce, sketch, show, compute, modify).

Analysis: identify elements, relationships, and organizational principles included in communications (e.g., analyze, compare, criticize, inspect, select, contrast, outline).

Synthesis: assemble elements and parts to form a whole including unique communications and sets of abstract relations (e.g., arrange, compile, invent, generate, construct, organize).

Evaluation: judge the value of materials and methods for a given purpose using internal and external evidence (e.g., assess, criticize, estimate, discriminate, judge, summarize).

Affective

Receive: attend willingly to stimuli (e.g., accept, choose, locate, name, listen, show).

Respond: respond actively to stimuli (e.g., answer, complete, describe, present, report, specify).

Value: accept a set of values (e.g., adopt, agree, choose, differentiate, initiate, recommend).

Organize: establish relationships among values (e.g., classify, compile, construct, design, manufacture, produce).

Characterize: behave consistently according to a set of values (e.g., act, behave, contradict, declare, defend, integrate, profess).

Psychomotor

Reflex: exhibit involuntary action or response (e.g., blink, hiccup, sneeze, twitch).

Motor: display locomotive and manipulative skills (e.g., raise, run, stand, sit, walk).

Perception: use kinesthetic, visual, tactile, and coordinated actions (e.g., feel, hear, see, smell, sense, taste).

Physical: display strength, agility, dexterity, and endurance (e.g., catch, dance, draw, kick, jump, march, throw).

Nonverbal: use facial and bodily expression and movement (e.g., dramatize, exhibit, gesture, mimic, pantomime, perform).

Communication: use verbal and written communication (e.g., debate, declare, describe, narrate, relate, sing, tell, write).

Test Validity

Validity is the extent to which a test measures what it is supposed to measure. It is the most critical aspect of a test.

Major types of validity:

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

Content validity is the degree to which a test corresponds to the curriculum and accurately assesses attitudes, skills, and knowledge included in the training program. Usually it is determined by using expert judgment which involves a review by qualified, experienced experts to determine if the test is accurate, appropriate, and fair.

Criterion-related validity compares 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 test scores relate to the theory on which a test is based by measuring a theoretical construct. For example, a researcher might examine a personality test to determine if the personality typologies account for actual results.

Improving Test Validity

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.

Increase test validity by using the following procedures:

Clearly specify the instructional objectives.

Match items to these objectives.

Use items which reflect proportionately the amount of training time spent on those topics.

Have experts independently review the items.

Write clear items using vocabulary appropriate for the intended group of examinees.

Use clear, simple test instructions.

Increase content validity by interviewing trainees to determine if they felt the test was appropriate.

Analyze test scores to identify "poor" items such as those answered incorrectly by many students.

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. Develop tests that include enough items to reflect the amount and level of content in the curriculum.

Cognitive Level

Competence	Know	Comprehend	Apply	Analyze	Synthesize	Evaluate
1	10%	10%				10%
2	5%	5%	10%			
3		15%				
4				10%		
5	15%				10%	
Total	30%	15%	25%	10%	10%	10%

When commercial publishers develop standardized, norm-referenced tests, they identify the curriculum used by 80 percent or more of students at the grade level for which the test is intended and prepare test items that correspond to the objectives for that curriculum.

Reliability

"Reliability coefficient" is a generic term referring to various types of reliability measures. Since reliability measures have different meanings, describe the method used to determine reliability.

Reliability is the extent to which test results are consistent, stable, and free of error variance. Error variance results from chance differences (e.g., time of day when the test is administered) or the use of ambiguous, confusing terms.

A reliable test may not be valid. A yardstick only 35 inches long will measure consistently, but inaccurately, resulting in invalid data.

Reliability is affected by several factors:

Trainee response variation due to physiological or psychological factors (e.g., amount of sleep; motivation).

Changes in test content.

Testing conditions (e.g., temperature, noise, or apparatus functioning).

Observational errors (e.g., mistakes in recording or scoring).

Reliability increases with the length of the test and the spread or variance of scores.

Difficult items which result in excessive guessing reduce reliability.

If a test is designed to predict a particular criterion (predictive validity), validity is more important that reliability (e.g., tests of creativity).

All else being equal with tests that measure the same thing, use the test with the highest reliability.

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.

Types of reliability coefficients:

Stability (test-retest): correlation between two successive measurements with the same test. Too short a delay between testing sessions may result in a spuriously high effect due to recall or a spuriously low effect if the delay is too long.

Equivalence (alternate forms): the successive administration of two parallel forms of the same test. This is the best index of test reliability.

Internal Consistency

Split-half divides the test into two equivalent halves (odd vs. even numbered items) and the Spearman-Brown formula is used to estimate reliability.

Cronbach's alpha compares each item's variance to total test variance.

Rational Equivalence: The Kuder-Richardson formulas 20 and 21 provide relatively conservative estimates of the coefficient of equivalence. Formula 21 is less accurate, but simple to compute.

Decision-consistency: Use decision-consistency with criterion-referenced tests. It is the average of squared deviations from the established mastery

Item Analysis

Item discrimination compares performance on each item for trainees with the highest and lowest test scores. Most frequently the highest and lowest 27 percent are compared.

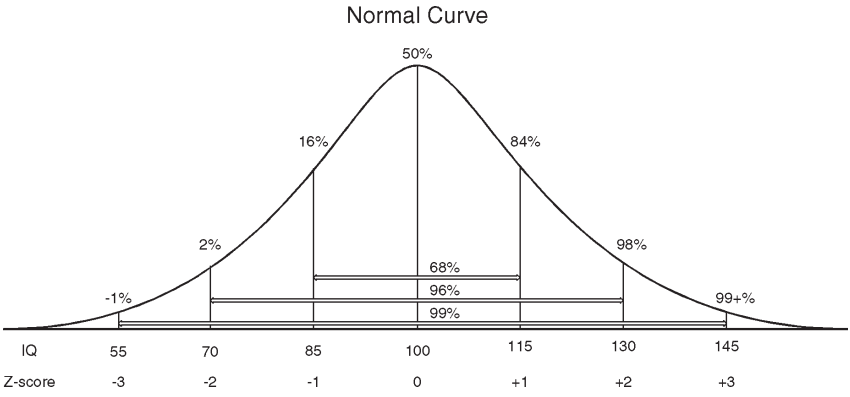
Item difficulty is the proportion of trainees responding correctly to an item. The statistic ranges from .00 to 1.00. The higher the proportion, the easier the item. Theoretically, the ideal level of difficulty is .50 if the effects of guessing are eliminated.

Item discrimination is the extent to which items distinguish among trainees in the high and low groups. These scores range from -1.00 to +1.00 with an ideal score of +1.00.

As a rule of thumb, revise or eliminate items with a discrimination index from 0 to + or -.30. Items with a negative discrimination favor trainees in the low group and should be revised. Retain items with discrimination indices ranging from + .30 to + 1.00.

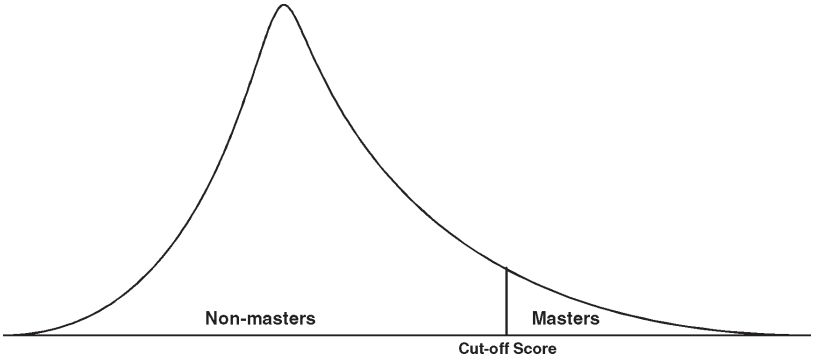
Item difficulty and discrimination are important concepts in developing standardized tests, including intelligence tests such as the *Wechsler Intelligence Scale for Children* (WISC). Scores on these tests are expected to approximate a normal curve distribution. On the WISC, for example, which has a mean of 100 and a standard deviation of 15, a score of 115 would fall one standard deviation above the mean at the 84th percentile.

Figure 1



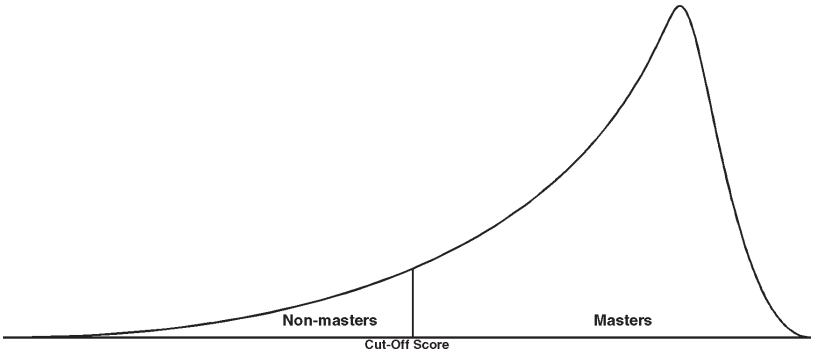
Mastery tests establish cutoff scores at a pre-established level that depends on the purpose of the test. Figure 2 illustrates a situation in which mastery was set a point that would fail most persons who took the test. This criterion level is appropriate for highly selective programs that admit few applicants, such as medical school or professional basketball.

Figure 2 High Cut-off Score - Low Level of Mastery



Many programs, particularly competency-based training, are designed to have most of the trainees achieve mastery. The distribution of posttest scores on these tests would resemble the curve shown in Figure 3.

Figure 3 Low Cut-off Score - High Level of Mastery



Pretest-Posttest Score Analysis

If most trainees have high pretest scores and high posttest scores, it is difficult to assess training effects. However, if most of the trainees do poorly on the pretest and master the posttest, conclude that training successfully improved performance. Ideally, trainees should answer all pretest questions incorrectly and all posttest questions correctly.

Determine how many trainees answered each item correctly on the pretest and posttest. If a large percentage — more than 50% — answer a pretest item correctly, examine the item, the related objective, and corresponding section of curriculum. Probably one or more of these should be changed.

If more trainees answer an item correctly on the pretest than on the posttest, it should be revised or eliminated. It may be poorly written or include material improperly presented during training.

Examine the percentage of trainees who select each distracter. Patterns of incorrect responses can help trainers detect misunderstandings, ambiguity, lack of knowledge, guessing, or even incidents of miskeying on the master correct answer sheet.

Conclusion

Test preparation is a major responsibility of trainers. Too often, qualified staff who prepare lessons carefully and teach conscientiously use inadequate tests that do not validly reflect the true level of trainee achievement. Testing has become increasingly important as state and federal agencies monitor the outcomes of training and educational programs. Curriculum developers sould devote greater attention to developing good tests that document program success and provide data improve instruction.

References

Ahmann, J. S., & Glock, M. D. (1971). *Evaluating pupil growth: Principles of tests and measurements*. Boston: Allyn and Bacon.

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

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

Bloom, B. S., Englehart, M. D., Frost, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain*. New York: David McKay.

Educational Testing Service. (1959). *Making the classroom test: A guide for trainers*. Princeton, NJ: ETS.

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

Harrow, A. J. (1970). *A taxonomy of the psychomotor domain: A guide for developing behavioral objectives*. New York: McKay.

Gay, L. R. (1980). *Educational evaluation and measurement: Competencies for analysis and application*. Columbus, OH: Charles E. Merrill.

Kibler, R. J., Barker, L. L, & Miles, D. T. (1979). *Behavioral objectives and instruction*. Boston, MS: Allyn and Bacon.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1956). *Taxonomy of educational objectives: Handbook II: Affective domain*. New York: David McKay.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.). *Educational Measurement* (3rd ed.). Washington, D.C.: American Council on Education, 335-366.

Popham, W. J. (1990). *Modern educational measurement: A practitioner's perspective*. Englewood Cliffs, NJ: Prentice-Hall.

Sax, G. (1974). *Principles of educational measurement and evaluation*. Belmont, CA: Wadsworth Publishing.

Thorndike, R. L., & Hagen, E. (1969). *Measurement and evaluation in psychology and education* (3rd ed.). New York: John Wiley.

Tuckman, B. W. (1975). *Measuring educational outcomes: Fundamentals of testing*. New York: Harcourt Brace Jovanovich.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.). *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 81-129.

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)