DOCUMENT RESUME

ED 304 450 TM 012 681

AUTHOR Pigge, Fred L.; Marso, Ronald N.
TITLE Supervisors Agenda: Identifying and Alleviating Teachers' Test Construction Errors.
PUB DATE Nov 88
NOTE 50p.; Paper presented at the Annual Conference of the Ohio Association for Supervision and Curriculum Development (Columbus, OH, November 3-4, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Administrator Attitudes; *Administrator Role; Classroom Techniques; Elementary Secondary Education; Public Schools; *Supervisors; Teacher Attitudes; Teacher Effectiveness; *Teacher Made Tests; *Test Construction; Testing Problems
IDENTIFIERS Ohio; *Teacher Competencies

ABSTRACT

Selected findings are presented from a study of classroom teachers' testing needs, testing proficiencies, testing practices, and testing resources in the public schools in Ohio. Focus was on determining how principals and supervisors can assist: (1) teachers in identifying and alleviating the most common test construction errors found on teacher-made tests; and (2) Ohio schools in providing resources to better support teachers' testing responsibilities. A sample of 586 school supervisors and principals and 326 classroom teachers completed administrators' and teachers' versions, respectively, of an assessment of needs instrument for 45 identified competencies and of perceived proficiencies in those competencies. Administrators and teachers also rated the availability of resources for testing and guidelines. Teachers further described their testing practices and provided a total of 175 samples of teacher-made tests. Teachers scheduled teacher-made tests frequently and used a variety of item types in making the tests; most teachers constructed their own items. Competency needs for testing were rated higher than beginning teachers' testing proficiencies by administrators. Teacher proficiency was rated highest by teachers and lowest by supervisors, with principals in the middle. Administrators and principals generally agreed on proficiency needs. Available resources appeared inadequate to support testing responsibilities. Teacher-made tests contained errors in format or construction. Guidelines and item type error formats for identifying and alleviating test construction errors are presented, with test examples. Seven tables present study data. (SLD)

Supervisors Agenda: Identifying and Alleviating Teachers' Test Construction Errors

Fred L. Pigge and Ronald N. Marso

College of Education and Allied Professions

Bowling Green State University

Bowling Green, Ohio 43403

A paper presented at the annual fall conference of the

Ohio Association for Supervision and Curriculum Development

Columbus, Ohio

November 3 & 4, 1988

Running Head: ALLEVIATING TEST ERRORS

Abstract

This paper presents selected findings from an assessment of classroom teachers' testing proficiencies by Ohio supervisors, principals, and teachers and from an analysis of actual teachers' testing proficiencies as displayed on samples of their teacher-made tests. Findings related to the availability of resources in Ohio schools to support teacher testing responsibilities, teachers' test planning and construction proficiencies, the nature and frequency of test construction errors found on teacher-made tests, and descriptions of the cognitive functioning levels of teacher-made tests are presented. The focus of the paper is upon how principals and supervisors can assist teachers in identifying and alleviating the most common test construction errors found on teacher-made tests and upon how principals and supervisors can assist Ohio schools in providing resources to better support teachers' testing responsibilities.

Supervisors Agenda: Identifying and Alleviating Teachers' Test Construction Errors

It is commonly understood that the professional literature and the professional advice given to teachers about the development and use of teacher-made tests are derived from a consensus of professional judgment rather than from knowledge acquired from research (Dwyer, 1982). For example, Gullickson (1984) states that we simply do not know how classroom tests are being used, and questions like, are they being used effectively, are even further from our present knowledge. Additionally, Stiggins and Bridgeford (1985) maintain that we have relatively little knowledge of what resources are available in the public schools to support classroom teachers' testing responsibilities.

Some recent research literature in the field of teacher testing, although limited to teacher reports rather than direct analyses of teachers' tests or direct observations of teachers' testing practices, has provided some understanding of classroom teachers' attitudes about testing and of classroom testing practices. This teacher self-report research literature suggests: that teachers have positive attitudes about the impact of testing on student learning and do schedule classroom tests frequently (Gullickson, 1984), that testing procedures vary somewhat by grade level and subject area (Stiggins & Bridgeford, 1985), that teachers place a heavy emphasis on informal observations and

assessments of teacher-pupil interactions as well as on formal tests (Salmon-Cox, 1981), that teachers seldom complete even relatively simple statistical analyses of the results of their testing efforts (Gullickson & Ellwein, 1985), and that teachers are more likely to design tests around curriculum guide objectives rather than through use of a test specification table and that most teachers use percentage correct grading and scoring procedures (Rogers, 1985).

The professional literature provides few studies of teachers' test construction skills as revealed through direct analyses of teacher-made tests. Both Billeh (1974) and Black (1980) reported studies involving the direct assessments of teacher-made tests; however, these studies were limited to the analyses of the cognitive functioning levels of science tests. They found that the cognitive demands of the science tests varied by field of specialization but not by the amount of training received by the teachers who had constructed the tests. The biology and chemistry tests contained proportionately more knowledge level test items than did the physics tests. In a more extensive analysis of teacher-made tests, Flemming and Chambers (1983) assessed 8,800 items contained in a sample of 342 tests. They found that short response, including fill-in-the-blank formats, followed by matching exercises were the most frequently used item types and that essay type items were the least frequently used type of item.

Their assessment of item cognitive functioning levels of the tests indicated that the junior high level tests contained proportionately the most knowledge level items (94%); whereas both the elementary and secondary level tests were comprised of about 69% knowledge items. These average percentages of knowledge level items by grade level, however, were found to be misleading, for when the tests were classified by subject area, it was found that the items functioning beyond the knowledge level were located almost exclusively on the math and science subject area tests. Additionally, these researchers found frequent format errors on the tests including items not being numbered consecutively, lack of directions for some or all exercises, illegible and/or handwritten text, and grammatical, spelling, or punctuation errors.

Purpose

The purpose of this paper was to present selected findings from a broader investigation of classroom teachers' testing needs, testing proficiencies, testing practices, and testing resources in the public schools of Ohio. Full details of the findings from the larger investigation are available elsewhere; the goal of this paper is to select and present findings from the larger study which appear to have direct implications for supervisors of teachers who wish to better understand teacher testing practices and to better assist teachers improve their teacher-made tests.

Illustrative of the types of questions addressed in this paper are: a) How many teacher-made tests per academic year does a typical classroom teacher construct? What percentage of the items on these tests are constructed by the teachers? What types of test items are most commonly used on these tests? b) Do teachers, principals, and supervisors perceive beginning teachers' test construction proficiencies to be adequate to meet classroom instructional needs? Do analyses of actual teacher-made tests confirm these perceptions? c) What types of test construction errors are most frequently made by classroom teachers? How can these errors be alleviated? d) At what cognitive levels are most teacher-made test items functioning? How can teachers improve the cognitive functioning levels of their tests? e) What resources are available in Ohio schools to support teachers' classroom testing responsibilities? What can be done to improve the availability of these resources in order to improve the quality of teacher-made tests?

The Subjects

The administrative subjects for this study consisted of 800 Ohio public school supervisors and principals randomly selected from the state directory of schools. The type of school system (city, exempted village, and county local), job assignment (principal and supervisor), and grade level assignment (elementary, middle, and secondary) classifications were used as strata in the

selection of the administrators. Responses to the assessment instrument sent to the selected 800 administrators after two follow-up contacts to nonrespondents resulted in usable responses from 586 (73%) administrators who identified themselves as supervisors (229), principals (313), and individuals in related (coordinators of curriculum or instruction, etc.) supervisory roles (44).

The teacher subjects were selected by "matching" the social security numbers of Bowling Green State University teacher-education graduates during the years of 1975 through 1985 with the social security numbers of full-time teachers certified by the Ohio State Department of Education for the 1985-86 academic year. This procedure resulted in the identification of 600 teachers from whom usable responses were obtained from 326 (54%). Only data obtained from teachers assigned to regular classroom instructional responsibilities were used for this report (specialized area teachers were excluded, e.g., art, music, special education, etc.).

Assessment Instrument

The assessment instrument consisted of 45 testing competencies located under four separate headings: a) working with teacher-made tests, b) using teacher-made test scores, c) working with purchased tests and scores in cumulative folders, and d) working with competency or mastery testing programs. The

respondents were directed to respond to the competency-mastery testing section only if their schools were involved in such programs. Both the administrator and teacher forms of the assessment instrument contained these four sections appearing in identical format. Each of the 45 competencies was responded to via two five-point Likert scales marked from high (5) to low (1) with headings for the administrators identified as "need of this competency to be a successful teacher in your school" and "average proficiency of your new teachers in this competency;" whereas the two Likert scales for the teachers' form were identified as "to be successful in your job, what is your need for this competency" and "an estimate of your classroom proficiency in this area."

In addition to the 45 testing competency items on the assessment instrument, both the administrators and the teachers were asked to report on the availability of 12 resources or guidelines to support teachers' testing responsibilities in their schools and were asked to assess via a Likert scale format the overall teachers' adequacy in tests and evaluation skills as compared to: a) knowledge of their subject area, b) proficiency in their other professional education competencies such as planning lessons, handling discipline, etc., and c) their overall competence or proficiency as educators. The teacher form of the assessment instrument also contained one additional section asking the teachers to report on seven testing preferences and practices,

such as how frequently they scheduled formal teacher-made tests and what types of test items they most frequently used in developing their classroom tests.

Teacher-Made Tests Sample

In addition to completion of the survey instrument, the 326 teachers were asked to enclose a copy of their most recently developed formal teacher-made test (not a quiz or a test from a spelling or a math class unless they were a math teacher) which resulted in the collection of 175 (54%) tests. These tests, regardless of grade level, when classified by subject area content consisted of 30 history/social studies, 36 science, 29 business education, 32 mathematics, 28 English, and 20 tests within nine other specializations with insufficient numbers to be included in distinct subject area categories.

The sample of 175 teacher-made tests included a total of 6529 test items and 455 item exercises. The test items within the example of tests were each classified independently by two judges using Bloom's taxonomy of six cognitive demand levels (knowledge, comprehension, application, analysis, synthesis, and evaluation). If the judges differed in their classification of an item or exercise, the item or exercise was reexamined until a consensus was reached. Each test and each test exercise was also examined for format and item construction errors. A test exercise was defined for this study as a group of items of a similar item type,

and item construction error criteria were selected from a review of several test construction textbooks designed for preservice (and inservice) education courses. A total of eight item type classifications (completion, essay, multiple-choice, etc.), 10 item format construction error criteria (does the test have complete directions? are item types grouped together? are the items numbered consecutively? etc.), and 66 item construction error criteria (incomplete stems, implausible alternates, specific determiners, etc.) were identified from these procedures and used in the assessment of the sample of teacher-made tests. An item construction error, if present, was recorded once per item exercise rather than for each time that particular error type may have occurred within an item exercise. In other words, regardless whether a construction error appeared only on one item or on several items within the same item exercise a tally of '1' was recorded for that particular error in order to provide a stable base of comparison across tests which varied in their number of test items.

Selected Findings

A. Nature of Teachers' Tests and Testing

1. Teachers schedule teacher-made formal tests (not including quizzes, spelling, etc.) frequently.

    a. The "average" teacher gives 54.1 formal tests during an academic year.

b. The typical teacher in a typical course gave a formal exam about every two weeks.

2. Teachers use a variety of item types in making tests.

a. The average teacher-made test consisted of 2.6 different item-type exercises.

b. The average teacher-made test contained 37.9 items.

c. The "average" teacher used the following percentages of item types to total items used: 20% multiple choice, 19% matching, 17% short response, 14% true-false, 14% problems, 8% completion, 6% interpretive exercises, 1% essay.

3. Teachers obtain items from more than one source, but most reported that they construct their own test items.

a. One-half the teachers constructed 75% or more of their items.

b. About 37% of the teachers reported constructing almost all of the items used on their tests.

c. Secondary teachers wrote more of their own test items than did elementary teachers.

B. Assessment of Classroom Testing Needs and Teachers' Testing Proficiencies (see Tables 1, 2, & 3).

1. Testing competency needs for success in the classroom are rated higher than beginning teachers' testing proficiencies.

a. Principals and supervisors rated all classroom testing competency needs higher than they rated beginning teachers' proficiencies in these competencies.

b. Teachers rated some but not all of their classroom testing competency needs higher than their testing proficiencies.

c. Regarding the three groups and their rating of teachers' testing proficiencies, the teachers rated their own level of proficiency highest, principals' ratings were in the middle, and supervisors' ratings were the lowest.

2. Teachers', principals', and supervisors' ratings of classroom testing needs and teachers' testing proficiencies correlate positive and high (e.g., the rater groups agree on which needs and proficiencies are highest, those in the middle, and lowest with relatively few exceptions).

3. Teachers, principals, and supervisors each rated the adequacy of teachers' testing and evaluation skills below average when they were asked to compare these skills with teachers' subject area knowledge, teachers' other professional education skills, and teachers' overall educational proficiency or competence.

4. Teachers', principals', and supervisors' ratings of teachers' test item construction proficiencies correlate moderately high but in the negative direction with teachers' item construction skills as displayed on their teacher-made tests (e.g., Specific item writing skills rated high were found to be the most error prone item exercises on the teachers' tests, and those rated low were the least error prone on the tests.).

C. Availability of Testing Resources in Ohio Public Schools (see Table 4) for the Support of Teachers' Testing Responsibilities.

1. Resources available in the public schools of Ohio appear to be inadequate to support teachers' testing responsibilities.

a. Just 50% of the teachers reported that test duplication and typing assistance are available to them.

b. Just 7 to 14% of the teachers reported the availability of grade assignment/deriving term grade guidelines.

c. Many (beginning teachers in particular) teachers reported that textbook instructor manuals (often with objectives and test items) were not available.

d. Just 16 to 26% of the teachers reported the availability of computer test scoring or related testing computer services.

2. Principals, supervisors, and teachers generally concurred on what resources were or were not available in the schools.

3. Suburban teachers reported more resources available to them than did rural teachers, and urban teachers reported the availability of even fewer resources than did either of the other two groups of teachers.

D. Cognitive Functioning Levels of Teacher-Made Tests (see Table 5).

1. Teachers, principals, and supervisors each rated to a high degree the teachers' need to construct test items measuring critical thinking type processes (upper cognitive levels).

2. Principals and supervisors rated teachers' proficiency in writing higher cognitive functioning test items very low; whereas teachers' rated this proficiency about average among their skills in working with teacher-made test.

3. The analyses of the 175 actual teacher-made tests revealed that 72% of the items thereon measured at the knowledge level, 11% at the comprehension level, 15% at

the application level, and about 1% at the levels beyond application (analysis, synthesis, evaluation).

a. The percent of knowledge items found on the teachers' tests varied with grade level and subject content area.

b. Secondary teachers wrote proportionally fewer knowledge items than did elementary teachers.

c. Social studies test items were 98% knowledge level, and most test items not in the math or science content areas were almost exclusively written at the knowledge level. Over one-half of all items functioning beyond the knowledge level were found to be on the math and science tests.

E. Test Format Errors Found on Teacher-made Test (see Table 6).

1. Teachers (rated second highest of 17 competencies) and administrators (rated among top one-third competencies) rated teachers' test format writing skills as high.

2. The direct analysis of the teachers' tests revealed an average of 1.6 test format construction errors per test (among the top one-third in frequency relative to the other types of construction errors identified on the teacher-made tests).

3. The most commonly identified types of test format errors found on the 175 teacher-made tests were (from highest

to lowest frequency with percentage of all test format errors identified in parentheses):

a. Absence of directions (29%), found on 82 of the tests.

b. Answer procedures not clear (22%), found on 61 of the tests.

c. Items not consecutively numbered (17%), found on 47 of the tests.

d. Inadequate margins (8%), found on 22 of the tests.

e. Answer spaces not provided (7%), found on 21 of the tests.

F. Test Item Construction Errors Found on Teacher Tests (see Tables 6 & 7)

1. The matching exercises were by far the most error prone item type found on the teacher-made tests.

a. An average of 6.4 types of different errors were found on the average matching exercise.

b. The matching exercises accounted for 58% of all item errors found on the sample of teacher-made tests.

c. Both the administrators' and teachers' ratings of teachers' competency in the writing of matching exercise were highest among all item types; whereas

the direct analyses of the teacher-made tests revealed this to be the most error prone item type.

2. The completion item-type with an average item exercise error rate of 2.2, the essay item type 1.5, the true-false item type 1.0, the multiple-choice item type .8, the short response item type .7, problem item type .5, and the interpretive item type with average exercise error rate of .2 all were much less error prone as compared to the matching exercise.

Guidelines for Alleviating Common Test Construction Errors

Guidelines and item type formats for identifying and alleviating test construction errors are presented in this section of the paper. Item types with associated guidelines are presented with items found to be most error prone (matching exercises) to least error prone (interpretive exercises).

Matching Exercise Format

Directions: (3 parts)
Establish basis for match. → In the left-hand column below are descriptions of some late-nineteenth century American painters. For each description, choose the name of the person being described from the right-hand column

How to answer → and place the letter identifying it on the line preceding the number of the description.

Avoid elimination. → Each name in the right-hand column may be used once, more than once, or not at all.

Column Titles → Description of Painter	Name of Painter
(e) 11. A society portraitist, who emphasized depicting a subject's social position.	a. Mary Cassatt
(d) 12. A realistic painter of nature, especially known for paintings of the sea.	b. Thomas Eakins
(b) 13. A realistic painter of people, who depicted strong characterizations.	c. Winslow Homer
(a) 14. An impressionist in the style of Degas, who often painted mother and child themes.	d. John LaFarge
	e. John Sargent
	f. James Whistler

Responses lettered, arranged in logical order, and to right (or top)

Premises (longer) numbered consecutively with test and to left side

premises and responses are homogeneous (e.g., all painters) and unequal numbers

MATCHING EXERCISES

Problems	How to Handle
1. Elimination problem	1.a More responses (or premises) than premises 1.b Responses (and premises) are homogeneous 1.c Responses used once, more than once, not at all
2. Premises lack of clarity (basis of match not clear)	2.a Premises must be sufficiently long to be clear or complete interrogative sentences 2.b Basis for match spelled out in directions
3. Waste of testing time, undue student frustration	3.a Arrange complete exercise single page 3.b Place letter of response in blank to left of premise (not write out answers) 3.c Use no more premises (or responses) than 6 to 10 3.d Responses logically ordered (alphabetical or chronologically)
4. Cognitive demand range (knowledge and comprehension typically)	4.a Names, dates, places, etc. require only knowledge (simple recall) 4.b Classifications, original examples of, predicted consequences require comprehension

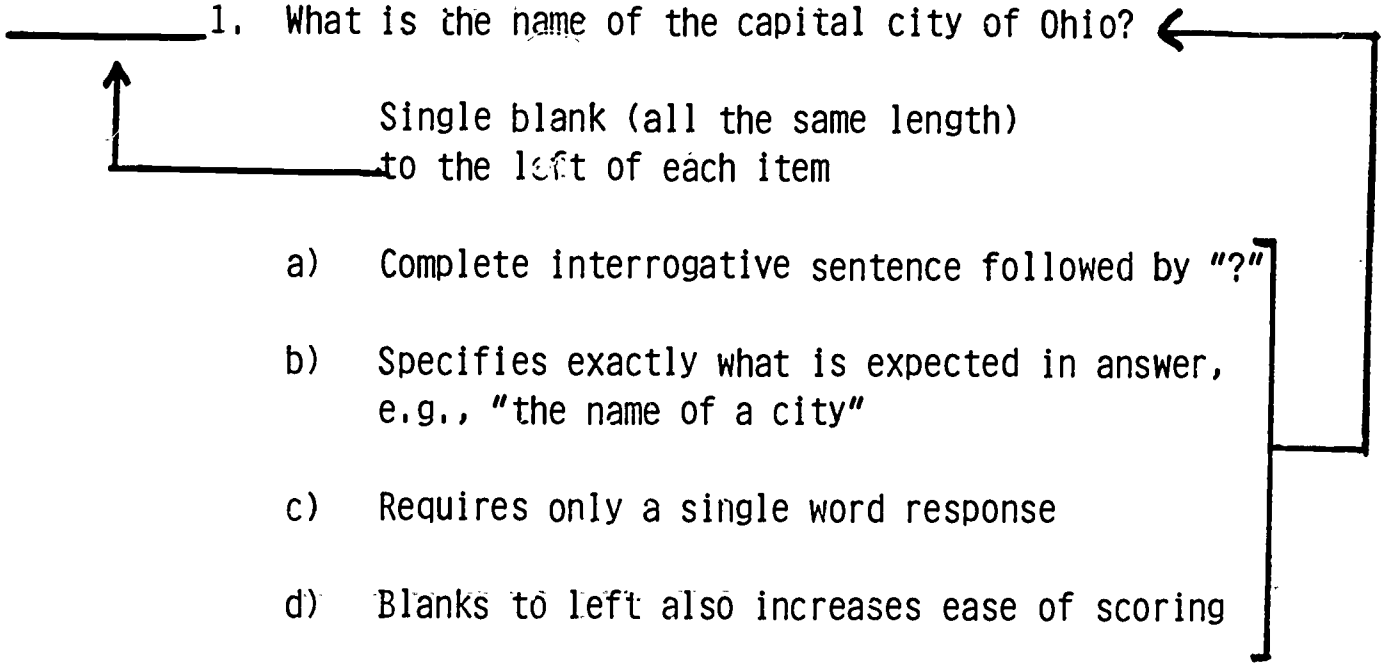
Problems	How to Handle
5. Inefficient format	5.a More lengthy phrases (premises) to left and responses to the right (or top) 5.b Premises numbered (consecutively within test) and responses lettered 5.c Answer blanks to left premises 5.d Columns (premises and responses) titled
6. Incomplete directions	6.a Spell out basis for match 6.b Indicate how and placement of answers 6.c "Responses may be used once, more than once, or not at all" (at least one of three each time)

Basic Concepts for Effective Matching

1. Avoid elimination
2. Homogeneous responses and premises
3. Basis for matching clear

COMPLETION (FILL-IN-BLANK) ITEM FORMAT

Directions: Read each question. Place the single word answer to the question in the blank to the left of the question.



The Relationship Among Constructed Response Items*

1. Completion: requires one word response
2. Short response: requires a phrase or sentence or two
3. Essay: requires typically paragraph or more response

*These definitions are arbitrary but this is a common distinction made among the three.

COMPLETION (FILL-IN-BLANK) ITEMS

1. Unclear question	1.a Write complete interrogative sentences
	1.b Place blank to the left (or right) of the question, e.g., do not place blanks in question statement
2. Ambiguous questions	2.a Write the question precisely so only one specific answer can be correct
	2.b Specify response expected, e.g., "Where was Jimmy Carter born? in hospital? in city? county? state? (ambiguous)
	2.c State the question so that a single word (only) is required.
	2.d Specify units/accuracy expected in the answer, e.g., in feet or yards.
	2.e State as clearly, concisely, appropriate vocabulary level, etc. as possible.

COMPLETION (FILL-IN-BLANK) ITEMS (cont.)

3. Presence of "unintended clues"	3.a Avoid clues such as number or length of blanks, grammatical clues, verb-object number (singular/plural) clues, etc. 3.b Do not give list of words/ answers to select from (this then becomes a matching exercise and would need to be designed accordingly).
4. Cognitive demand range (generally only knowledge)	4.a Avoid use of completion unless only simple recall, knowledge responses are desired 4.b Do ask main idea rather than "trivia", e.g., in what year was Jimmy Carter born vs on what day of the week. 4.c Avoid statements from the textbook with a word(s) (the blank) left out.

Basic Concepts for Effective Completion Items

1. Do not use textbook statements with words left out.
2. Use only complete interrogative questions with response (blank) to left.
3. Limit use to objectives where only knowledge (simple recall) is desired.

RESTRICTED ESSAY ITEM FORMAT

1. Directions ⟶
   a. Restrict task
   b. Alert students to answer all parts
   c. Warn about handling unrelated data

Directions: Read each question carefully and respond to and label your response to each part of the questions. The points assigned to each question are noted. Please confine your response to the space provided. Points will be taken off for incorrect and irrelevant data in your response.

2. Format ⟶
   a. Restricts student response
   b. Spells out expectations of student
   c. Indicates scoring weight
   d. Not limited to simple listing (knowledge) response

1. You find that your last examination had a KR_{21} reliability estimate of .57. This indicates that you must improve that test and are debating whether to add 20 more good completion items or to add 12 well-stated multiple-choice items in the additional fifteen minutes of testing time you have available. a) Select one of these strategies that will best improve your test, b) explain the pros and cons related to the choice of each option, and c) defend your choice (4 pts.).

RESTRICTED ESSAY ITEMS

1. Task too ambiguous for accurate scoring (lack of reliability)	1.a Restrict student response to knowledge acquired in unit (not measure general philosophy or general thinking ability) and restrict points assigned
	1.b Be fair in clearly spelling out expected student response, label each part of question
	1.c Write an answer-scoring key and model answer before final revision of the item
	1.d Take points off for irrelevant data (prevent bluffing)
	1.e Use point method for scoring, e.g., one point each main idea presented
2. Poor sampling of content	2.a Ask several brief rather than one or two very broad questions
	2.b Use essay to supplement objective items
	2.c Option questions are avoided or limited to options within content categories
	2.d Not request feeling or thinking, but evidence related to attaining unit content objectives
	2.e Avoid meaningless words like discuss, analyze, evaluate, compare and contrast

3. Cognitive range, complete range possible, purpose higher level	3.a Avoid tasks that require only simple listing, steps, names, places, ordering, etc. (knowledge, simple recall) 3.b Present problem requiring thoughtful solution, application of concepts and principles 3.c Use novel, hypothetical situations requiring critical thought
4. Unrealistically high scoring points	4.a Limit most responses from 2 to 5 points to not overweigh relative to objective items

Basic Concepts for Effective Essay

1. Present restricted task

2. Seek critical applications and thinking related to knowledge of unit (not general feelings, etc.)

3. Specify expectations in student response

4. Avoid simple listing tasks

ALTERNATE CHOICE (TRUE-FALSE) FORMAT

Directions: Read each question carefully and determine if the statement is true or false. If the statement is true circle 'T,' or if the statement is false circle the 'F' before the statement.

T F 1. The capital city of Ohio is Columbus.

a) 'T' and 'F' is typed to the left of each statement.

b) Circling a letter to the left provides ease in scoring and accuracy, e.g., less difficulty in scoring when answer is changed by student.

c) Concise, clear statement with simple sentence structure.

d) Statement must include only single idea which is clearly either true or false.

Typical Alternate Responses

1. True - false
2. Fact - opinion
3. Complete - incomplete sentence
4. Event - consequence
5. Solid - liquid
6. Acceptable - unacceptable

ALTERNATE CHOICE (TRUE-FALSE) ITEMS

Problems	How to Handle
1. Inappropriate content	1.a Content or statement must be such that it is unequivocally true or false (not it depends). 1.b Opinion statements are excluded or the opinion is attributed to a source.
2. Complex structure, dual ideas, negatives	2.a Limit statement to a single central, significant idea (so not part true and part false). 2.b Write concise statements with simple sentence (not compound or compl(x) structure. 2.c Negative sentences are not acceptable in true-false; they must be rewritten to positive statements and rekeyed.
3. Presents irrelevant barriers	3.a Avoid negative and double negative statements. 3.b Avoid lengthy, inappropriate vocabulary level, complex sentence structure
4. Time waste	4.a Have students circle T or F and not write out answer. 4.b Application of correction for guessing formula, correcting of false statements, etc. usually do not warrant extra time required.

Problems	How to Handle
5. Presents unintended clues	5.a Avoid "specific determiners" such as never, all, always, etc. 5.b Avoid length clues (true statements may tend to be longer) 5.c Avoid answer patterns, e.g., 50% true and 50% false, or true-false true-false sequences.
6. Cognitive demand range (generally knowledge and comprehension)	6.a Avoid statements taken directly from text (add "not" or change one word, etc.) as encourages poor study habits of simple recall. 6.b Convert to novel examples, predictions of outcomes, etc. to reach comprehension level cognitive demand 6.c Avoid questions on "trivia" and only names, dates, places.

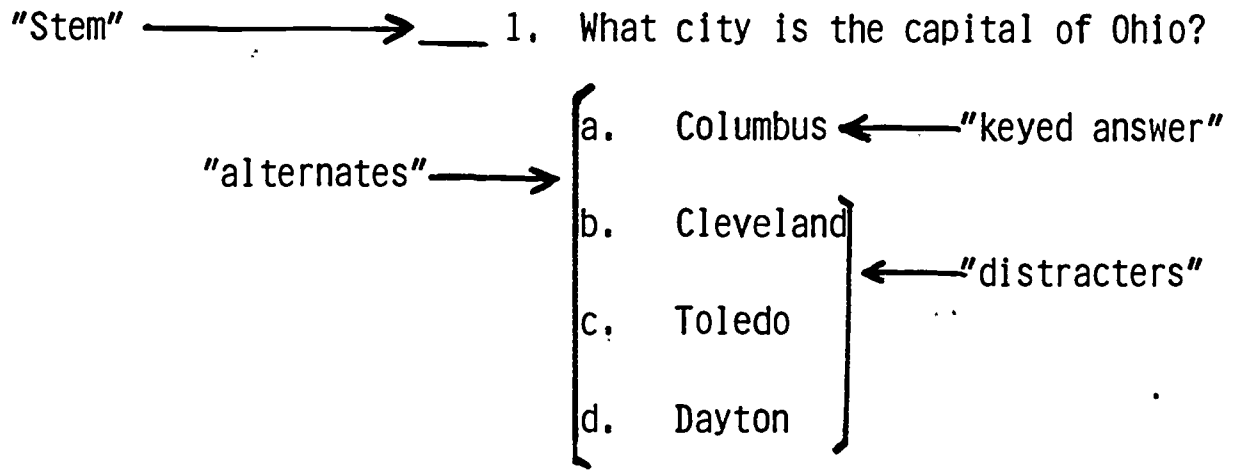
Basic Concepts for Effective True-False Items

1. Appropriate content (completely true or false).

2. Concise, single idea statements, avoiding "clues."

MULTIPLE CHOICE ITEM FORMAT

Directions: Choose the single best answer and place the letter of that answer in the blank to the left of that question.



Two forms for multiple choice items:

1. Correct response: only one correct answer.

2. Best response: all alternates correct but one clearly best (functions at higher, more desirable cognitive levels).

MULTIPLE CHOICE ITEMS

Problems	How to Handle
1. Incomplete stems	1.a Pose a clear problem or question in stem
	1.b Avoid one or two word stems
2. Distracters not feasible, plausible	2.a Make distracters homogeneous, feasible, logical
	2.b Distracters should reflect typical misconceptions
3. Undesirable "filler" distracters present	3.a Avoid "all above," "none of above" except when appropriate to problem posed
	3.b Do not use 'a' and 'b' but not 'c', etc. as distracters.
4. Presence of irrelevant barriers	4.a Write clear, concise stems and distracters.
	4.b Use positive statements, underline negative words when not avoided
	4.c Use simple and appropriate vocabulary
	4.d Avoid unnecessary repetition in alternates (place in stem)
	4.e Extraneous information, phrases avoided in stem
5. No single "best" answer	5.a Check to be sure single clearly best answer.
	5.b Avoid "overlapping" distracters.

MULTIPLE CHOICE ITEMS (cont)

Problems	How to Handle
6. Presence of unintended clues	6.a Avoid "grammatical clues," e.g., a/an.
	6.b Avoid "verbal association," e.g., words or phrases repeated in stem and answer only.
	6.c Avoid "specific determiners" in distracters, e.g., always, never, all.
	6.d Avoid "number clues," make stem and alternates singular or plural in structure, e.g., is/are.
	6.e Avoid length clue (e.g., answers longer than distracters).
	6.f Avoid overuse of positions of alternate as key (e.g., 25% correct a, b, c, & d).
	6.g Reduce guessing problem by using four alternates.
7. Unreadable format	7.a Place all distracters in column or row format.

MULTIPLE CHOICE ITEMS (cont)

Problems	How to Handle
8. Cognitive demand range, all levels	8.a Avoid textbook phrases or sentences (encourage simple recall study). 8.b Pose hypothetical situations, problems (what would happen if? How can this be corrected? fixed?). 8.c Present novel, new examples 8.d Require best judgment selection, based upon predictions, consequences, applications or principles, laws. 8.e Avoid or limit use of questions requiring only recall of names, places, dates, events, etc.

Basic Concepts for Effective Multiple-Choice

1. Use of feasible, logical, homogeneous distracters.

2. Avoiding "irrelevant barriers" to student knowledge.

3. Avoiding "unintended clues" so that even uninformed students get the answer.

4. Use of novel, hypothetical problems posed in stem demanding application, understanding, evaluative type responses.

SHORT RESPONSE ITEMS FORMAT

1. Specify response expected →

Directions: Answer each question with a few words or one sentence in the space provided. One point for each correct answer.

1. Why did Tom Sawyer become angry with the raft after the storm?

Provide appropriate response space. →

State questions requiring less than paragraph but more than one word.

2. What is likely to occur when a mixture of calcium granulates and sodium sulfate is mixed with hot water?

Problems	How to Handle
1. Unclear expectation	1.a Specify space, response nature, and scoring points.
2. Simple recall listing responses	2.a State questions where interpretations or understandings are required, e.g., do not ask questions only recalling names, dates, places, events, etc.
3. Ambiguous	3.a State concise, simple interrogative questions requiring phrases or a single sentence (not a word or paragraph)
4. Unrealistically high scoring points	4.a Assign usually single point or at most two (consider weight compared to other objective items on test)

PROBLEM SOLVING (NUMERICAL) ITEMS

Nature: Computation tasks in math, physics, business, etc.

Format: Variety of formats including narrative, pictorial, and numerical form.

Problems	How to Handle
1. Sampling content and time limitations	1.a Include wide range simple to complex items (rather than 1 or 2 complex).
	1.b Include both narrative (story problems) and numerical.
	1.c Use variety of other item types also to sample range of cognitive levels to include understanding concepts as well as calculation.
	1.d Group items measuring same processes together to save testing time, e.g., fractions together.
2. Minor errors and diagnosis concerns	2.a Ask students to show calculations to allow diagnosis and part-scores for correctness of procedures.

Problems	How to Handle
3. Complexity of task concerns	3.a Provide a range of simple to complex tasks. 3.b Simplify situations for clarity and to allow assessment of understanding. 3.c Sufficient space provided to complete calculations. 3.d Provide sufficient testing time so all items can be attempted.
4. Nonindependent items	4.a The correctness of one problem should not be dependent upon prior problem, e.g., not use answer to #16 for #17 calculations.
5. Cognitive range limitations	5.a Problem items should be accompanied by other item types to allow measurement at other than application level, e.g., other item types for recall, understanding, analysis, etc.
6. Precision of answer expected	6.a Be sure to specify accuracy and/or units of measure desired in answer, e.g., round to nearest one-tenth, square feet or square inches.

INTERPRETIVE EXERCISE FORMAT

Directions: →
a) Intro. to data
b) Denotes question no.
c) How to answer

Directions: For questions 11-15 please first read the information below related to an experiment on the transfer of genetic traits. After reading the information about the experiment choose the best answer and place the letter of that answer in the blank to the left of each question.

Data presented →
Data commonly, map, chart, graph, poem, cartoon, drawing, blueprint, passage, quotation, diagram, narrative problem, etc.

Experiment: In an experiment using fruit flies, a light bodied parent is crossed with a dark bodied parent. The offspring were all light bodied. Two light bodied offspring were then crossed, producing both light and dark bodied offspring in a ratio of 3 light to 1 dark. Using this information answer the following.

4 to 8 objective questions →

11. The parents of the second (P_2) cross are:
    a. hybrid
    b. pure
    c. heterozygous for light body
    d. none of these

12. The F generation in the second cross is
    a. hybrid
    b. pure
    c. homozygous for light body
    d. none of these

INTERPRETIVE EXERCISES

Problems	How to Handle
1. Presentation of data	1.a Data (problem, poem, map, etc.) is new to the student but related to unit and realistic.
	1.b Clear, concise, but sufficient to answer questions.
	1.c Pictorial data simplified, clear, accurately duplicated.
2. Objective items	2.a Item type should be objective and follows all construction guidelines.
	2.b Most measure at upper cognitive levels for interpretive exercise purpose.
	2.c Several questions should be used to compensate for time demands of data analysis, e.g., 4 to 8.
	2.d Questions must not be answerable without having read data presented in the exercise.
3. Cognitive range, purpose higher range	3.a Novel data, questions constructed for higher level required to accomplish purpose.

References

Black, T. R. (1980). An analysis of levels of thinking in Nigerian science teachers' examinations. Journal of Research in Science Teaching, 17, 301-306.

Billeh, V. Y. (1974). An analysis of teacher-made test items in light of the taxonomic objectives of education. Science Education, 58, 313-319.

Dwyer, C. A. (1982). Achievement testing. In H. E. Mitzel (Ed.), Encyclopedia of Educational Research, (4th ed., Vol. 1, pp. 13-22). New York: The Free Press.

Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. New Directions for Testing and Measurement, 19, 29-38.

Gullickson, A. R. (1984). Teacher perspectives of their instructional use of tests. Journal of Educational Research, 77, 244-248.

Gullickson, A. R., & Ellwein, M. C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. Educational Measurement: Issues and Practice, Spring, 15-18.

Rogers, B. G. (1985). Prospective teacher perceptions of how classroom teachers use evaluation methods: A qualitative research approach. Mid-western Educational Researcher, 613-20.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening. Phi Delta Kappan, 61, 631-634.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. Journal of Educational Measurement, 22, 271-286.

37/21

Table 1

Principals and Supervisors' Estimates of the Needs and Proficiencies of Beginning Teachers in 17 Test Development Competency Areas

		Means					
	Test Development Competencies	Need	Proficiency	Discrepancy	Rank*	t	p
1.	Writing multiple choice items	3.83	2.99	.84	12	19.53	.001
2.	Writing completion items	3.91	3.06	.85	11	19.75	.001
3.	Writing matching items	3.70	3.10	.60	15	13.73	.001
4.	Writing true/false items	3.51	2.99	.62	14	10.68	.001
5a.	Writing essay items	4.27	2.74	1.53	5.5	32.29	.001
5b.	Scoring essay items	4.35	2.67	1.68	3	36.06	.001
6.	Identifying good and poor items	4.34	2.83	1.51	7	35.15	.001
7.	Items harmony school/class goals	4.33	2.79	1.54	4	34.12	.001
8.	Stating clear/measurable objectives	4.40	2.87	1.53	5.5	35.26	.001
9.	Items measure higher thinking	4.45	2.55	1.90	1	38.29	.001
10.	Items measure true progress	4.50	2.78	1.72	2	38.39	.001
11.	Use less formal assessments	3.61	2.86	.75	13	15.95	.001
12.	Use observation assessments	4.02	2.96	1.06	9.5	24.14	.001
13.	Use sociometric type assessments	3.19	2.72	.47	16.5	10.70	.001
14.	Selecting items from manuals	3.60	3.13	.47	16.5	11.24	.001
15.	Attractive test format	4.08	3.02	1.06	9.5	24.46	.001
16.	Test coverage of text and class	4.51	3.19	1.32	8	32.18	.001
	Combined items totals	68.68	49.23				
	t-ratio		38.70				
	Probability level		.001				

*Rank ordered by magnitude of discrepancy

Table 2

Comparisons of Secondary and Elementary Principals', Supervisors', and Teachers' Ratings of Teachers' Test Construction and Planning Proficiencies*

	Secondary Means					Elementary Means				
	(1) Prin.	(2) Supr.	(3) Tchr.	F	Sch.**	(1) Prin.	(2) Supr.	(3) Tchr.	F	Sch.**
1.	3.06	2.95	3.71	31.35[a]	3>1,2	3.07	2.91	3.56	15.69[a]	3>1,2
2.	3.12	2.92	3.84	34.15[a]	3>1,2	3.15	3.03	3.53	7.87[a]	3>1,2
3.	3.15	3.02	3.92	44.47[a]	3>1,2	3.16	2.92	3.62	13.38[a]	3>1,2
4.	3.01	2.84	3.56	20.99[a]	3>1,2	3.11	2.99	3.49	5.78[b]	3>1,2
5.	2.87	2.47	3.67	36.18[a]	3>1,2	2.86	2.74	3.16	3.65[c]	3>2
6.	2.83	2.45	3.45	23.47[a]	3>1,2	2.76	2.55	2.84	2.07[d]	---
7.	2.90	2.67	3.85	47.59[a]	3>1,2	2.98	2.78	3.51	13.54[a]	3>1,2
8.	2.84	2.71	3.78	36.19[a]	3>1,2	2.93	2.76	3.57	14.90[a]	3>1,2
9.	2.95	2.85	3.63	19.39[a]	3>1,2	3.01	2.58	3.40	14.59[a]	3>1,2
10.	2.67	2.44	3.86	61.03[a]	3>1,2	2.67	2.47	3.27	11.48[a]	3>1,2
11.	2.81	2.56	3.68	38.83[a]	3>1,2	2.98	2.73	3.43	11.48[a]	3>1,2
12.	2.90	2.72	3.13	4.67[c]	3>2	2.97	2.85	3.27	4.16[c]	3>1,2
13.	3.00	2.90	3.44	9.68[a]	3>1,2	3.05	2.99	3.67	12.28[a]	3>1,2
14.	2.75	2.75	3.13	5.92[b]	3>1,2	2.74	2.78	3.27	8.90[a]	3>1,2
15.	3.09	3.16	3.80	23.44[a]	3>1,2	3.25	3.19	3.51	3.36[a]	3>1,2
16.	3.12	2.97	4.03	46.41[a]	3>1,2	3.07	3.00	3.60	9.01[a]	3>1,2
17.	3.29	3.18	4.35	58.36[a]	3>1,2	3.24	3.18	3.76	8.21[a]	3>1,2
++	50.36	47.31	63.14	71.39[a]	3>1,2	51.06	48.27	58.15	14.30[a]	3>1,2

* See Table 1 for description of competencies 1-17

** Alpha = .10 for these Scheffe post hoc pair-wise mean comparisons; read 1 = principals, 2 = supervisors, 3 = teachers; 3 > 1,2 reads teachers rated this proficiency higher than principals and supervisors, differences between principals and supervisors were not different

++ Totals all items combined

a = p's < .001 b = p's < .01 c = p's < .05 d's = p > .05

Table 3

Comparison of Principals' and Supervisors' Rating Means for Beginning Teachers' Testing and Nontesting Competencies/Skills

Relative Proficiency Rating Items*	Means			t**	p
	Principal	Supervisor	Combined		
1. Relative to knowledge of their subject areas, beginning teachers' test and evaluation competencies/skills are...	3.03	2.87	2.95	2.47	.014
2. Relative to their other professional education competencies, such as planning, discipline, etc., beginning teachers' test and evaluation competencies/skills are...	2.96	2.81	2.89	2.34	.020
3. Relative to their overall competence as educators, beginning teachers' test and evaluation competencies/skills are...	2.93	2.73	2.84	3.34	.001

*Ratings were recorded via a five point Likert-type scale, 5 (well above average), 4 (somewhat above average), 3 (about average), 2 (somewhat below average), and 1 (much below average)

**Ratios for t comparisons between the principals' and supervisors' rating means

Table 4

Teacher Responses to Availability of Testing Resources in Schools to Support Teachers' Testing Responsibilities

Resources/Guidelines	Availability		
	% Yes	% No	% In some subjects or sometimes
1. Typing and duplication assistance in preparing tests.	50	30	19
2. Convenient access to individual student records, tests, etc.	91	3	7
3. Counselor or other school staff to assist in interpreting class or individual standardized test results.	72	12	15
4. Curriculum guides with stated objectives for units of instruction.	87	4	9
5. Instructor manuals which provide you with questions for tests.	71	9	20
6. School or department guidelines on how many A's, B's, C's etc. to assign to a typical class at the end of the term.	7	88	5
7. School or department guidelines on relative weighting of the final term test or other scores in deriving final term grades.	45	49	6
8. School or department guidelines on how many scores or tests are required in deriving a term final grade.	14	80	6
9. Computer test scoring service for teacher-made tests.	22	71	8
10. Computer analysis of student responses to test questions.	16	72	11
11. Computer grade book record keeping for your classes.	26	57	17
12. Computer programs for generating tests for your classes.	22	57	20

Table 5

Judged Item Cognitive Functioning Levels by Item Type

Item Type	No. Items	% Beyond Knowledge	Number of Items Found at Each Cognitive Level					
			Knowl.	Compr.	Applic.	Analysis	Synthesis	Eval.
Completion	549	2	540	9	0	0	0	0
Matching	1261	8	1159	102	0	0	0	0
True/False	935	20	751	175	0	9	0	0
Multiple-Choice	1317	15	1123	7	112	73	2	0
Essay	64	53	30	22	6	1	1	4
Problems	896	96	35	59	798	4	0	0
Interpretive	362	35	199	118	40	4	0	1
Short Response	1093	24	830	235	28	0	0	0
Unclassified	52	46	28	23	0	0	1	0
Totals	6529		4695	750	984	91	4	5
Percent of total items at each level			72%	11%	15%	1%	.001%	.001%

Table 6

Summary of Format and Item Type Construction Error Frequencies

Test Item Construction Error Summary

	No. Items Reviewed	% Total Items Reviewed	No. of Exercises	No. Errors Present*	Mean Errors Per Exercise
Item Type Errors					
1. Matching	1261	19	78	496	6.4
2. Completion	549	8	48	106	2.2
3. Essay	64	1	22	34	1.5
4. True/False	935	14	69	71	1.0
5. Multiple-Choice	1317	20	65	53	.8
6. Short Response	1093	17	89	61	.7
7. Problems	896	14	54	26	.5
8. Interpretive Exercise	362	6	30	6	.2
9. Unclassified	52	1	6	-	-
Subtotals	6529	99	455	853	1.9

Test Format Construction Errors

Test Format Errors	No. Tests** Errors Present	% of Total
1. Absence of directions	82	29
2. Answering procedures unclear	61	22
3. Items not consecutively numbered	47	17
4. Adequate margins	22	8
5. Answer spaces not provided	21	7
6. Space between items not provided	12	4
7. Nonindependent items	11	4
8. Different weighting of objective items	8	3
9. Items not arranged most to least time demanding	7	2
10. Similar item types not grouped together	6	2
	281	100

*Each specific item type construction error (see Table 7 for listing of the specific error types and frequency) was tallied only once if present in an exercise (i.e., an error may have occurred several times or once in an exercise but in either case only a single tally was used so that tests and exercises could be compared regardless of the number of individual items appearing in a test or exercise).

**There were 175 individual tests but some tests had more than one error.

Table 7

Frequency and Nature of Item Construction Errors Found on Each Item Exercise

Construction Error	N	%*
a. Completion Item Type		
Not complete interrogative sentence	32	30
Blanks in statements	31	29
Textbook statements with words left out	18	17
More than single blank in statement	12	11
Question allows more than single answer	6	6
Blank number clue	4	1
Blank length clue	1	1
Requests trivia versus significant idea	1	1
Unstated degree of precision	1	1
Lengthy, unnecessary words or phrases	0	0
	106	100
b. True-False		
Required to write response, time waste	20	28
Statements contain more than single idea	16	23
Negative statements used	15	21
Presence of specific determiner	8	11
Statement not question, give away item	6	8
Needless phrases present, too lengthy	4	6
Imprecise statement, not always true or false	1	2
Presence of length clue	1	1
Opinion not attributed to source	0	0
	71	100
c. Essay Exercises		
Response expectations unclear, not labeled, etc.	14	41
Scoring points not realistically limited	7	21
Optional questions provided	5	15
Restricted question not provided	3	9
Ambiguous words used	2	6
Opinion or feelings requested	2	6
Question limited to simple listing response	1	2
	34	100
d. Problem Exercises		
Items not sample understanding concepts, only calculations	20	77
Not range of easy to difficult problems	3	12
Degree of accuracy not requested	2	8
Nonindependent items	1	4
Use of objective items when calculation preferable	0	0
	26	100

*Each specific item type construction error was tallied only once if present in an exercise (i.e., an error may have occurred several times or once in an exercise but in either case only a single tally was used so that tests and exercises could be compared regardless of the number of individual items appearing in a test or exercise), the percentage refers to percent of this error type to all errors found on all exercises of this type.

(table continues)

Construction Error	N	%*
e. Matching Item Type		
Columns not titled	71	14
Use once, more than once, or not all not in directions to prevent elimination	69	14
Response column not ordered	60	12
Directions not specify basis for match	55	11
Answering procedure not specified	52	10
Elimination due to equal numbers	46	9
Column(s) exceed 10 items	39	8
Materials not homogeneous	38	8
Premise not to left side	37	7
Numbers not to left and letters to right	13	3
Exercise not contained on single page	7	2
Requires responses to be written out	6	1
Insufficient information in premises	3	1
	496	100
g. Interpretive Exercises		
Objective response form not used	6	100
Can be answered without data presented	0	0
Errors present in response items	0	0
Data presented unclear	0	0
	6	100

Construction Error	N	%*
f. Multiple Choice		
Alternates not in column(s) or rows	21	40
Incomplete stems	12	23
Negative words not emphasized or avoided	9	17
"All or none above" not appropriately used	5	9
Needless repetition in alternates	2	4
Presence of specific determiners in alternates	2	4
Verbal associations between alternate and stem	1	1
Alternates overlap	1	1
Needless phrases used	0	0
Grammatical clues	0	0
Distractors implausible	0	0
Length clues	0	0
a and c, but not b, etc. used	0	0
	53	100
h. Short Response		
Item requires only listing	51	84
Response expectations ambiguous, not specified	7	11
Unrealistically high scoring values assigned	3	5
	61	100