DOCUMENT RESUME

ED 252 559 TM 850 028

AUTHOR Herman, Joan
TITLE Guidelines for Developing Diagnostic Tests. Methodology Project.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 84
GRANT NIE-G-84-0112-P1
NOTE 27p.
PUB TYPE Reports - Descriptive (141) -- Guides - Classroom Use - Guides (For Teachers) (052)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Diagnostic Tests; Elementary Secondary Education; *Skill Analysis; Student Evaluation; Teacher Made Tests; *Test Construction; Test Format; *Test Items; Test Reliability; Test Validity
IDENTIFIERS *Domain Referenced Tests; *Test Specifications

ABSTRACT

Diagnostic testing can provide specific information about student skills as a decision-making aid to teachers in prescribing instruction, identifying needs for remediation, determining effective instructional materials and methods, and ultimately, improving student learning. Diagnostic testing, as viewed here, includes individual and group assessment of students' skills in specified cognitive domains. A methodology is presented for designing diagnostic tests which assess the extent of student learning and are sensitive to sources of difficulty within a skill or context area. This 5 step methodology for diagnostic test development includes: (1) Developing a skill blueprint including a general description of the objective or skill, a sample item, content limits, and response limits; (2) Specifying the skill map including sub-skills or simpler contexts which students should master enroute to the desired skill under assessment; (3) Formulating test items that match specifications and follow conventions for sound item-writing; (4) Reviewing test items to insure match to specifications and technical quality; and (5) Field testing the items and revising to insure that the test is appropriate for the intended student population and structured to provide meaningful and reliable diagnostic information. (Author)

DELIVERABLE - NOVEMBER 1984

METHODOLOGY PROJECT

Guidelines for Developing Diagnostic Tests

Joan Herman

Study Director

Grant Number

NIE-G-84-0112, P1

Center for the Study of Evaluation

Graduate School of Education

University of California, Los Angeles

The project presented or reported herein was supported pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

GUIDELINES FOR DEVELOPING DIAGNOSTIC TESTS

Diagnostic testing can serve a variety of important roles in routine classroom practice which can help teachers to enhance their instructional effectiveness and improve student learning. By providing specific information about the entrance skills that students have and have not acquired, the types of tasks and subtasks they have mastered, and the nature of their errors and misconceptions, diagnostic tests are a useful tool in on-going instructional planning. They can be used:

- o at the beginning of the school year, to identify individual, group and/or class needs in order to prescribe appropriate instruction;
- o during the school year, to assess areas of instruction where individuals or groups of students are having difficulty and to identify specific needs for remediation;
- o throughout the school year, to identify areas where instructional materials and methods were effective and those which are in need of modification.

This view of diagnostic testing is both broader and more narrow than common definitions. First, it broadens the definition of diagnostic testing to include all tests which provide systematic information about what skills students have and have not acquired. Second, it moves beyond individual assessment to encompass both tests which can be used to make instructional decisions about individual students and tests which can be used to guide instruction for groups of students. It is narrow, however, in that it focuses on assessment of academic achievement to identify student strengths and weaknesses and does not consider the range of other relevant, non-cognitive factors which may elucidate the reasons for their performance. This latter focus is not meant to underestimate the significance of other factors in students' learning and instruction nor their importance in designing effective educational treatments. Diagnosis and prescription for individual students which ignores student affect,

motivation, nutrition, and vision, to name a few potentially relevant factors, is likely to be found wanting, and these factors must, of course, be included in a comprehensive diagnostic system. Nonetheless, a thorough understanding of what students can and cannot do, what skills they have and have not acquired, and of where there may be gaps in their learning, is at the heart of a sound diagnostic-prescriptive approach.

This paper provides a methodology for designing diagnostic tests which systematically assess the extent of student learning and seek to locate, where appropriate, sources of difficulty within particular skill or content areas. The approach is keyed to a teacher's or a curriculum's instructional intentions and considers students' status with respect to those intentions. That is, it starts with specific curriculum goals and objectives, creates a potential learning map by analyzing the subtasks, competencies and/or component skills that are necessary to the achievement of the desired objectives, and builds a test to chart students' progress with regard to the map. The map not only provides the means for diagnosing student difficulties, but also helps to clarify instructional intentions and to target instructional activities. The result is instruction which systematically teaches students necessary pre-requisites and builds their skills to desired levels. Under these conditions, the assessment is tied directly to the instructional context and its instructional implications are clear.

How does one accomplish building such a test? The following five steps can guide the test development process:

1. Develop a blueprint of the skill or content area you want to diagnose, i.e., clarify the nature of the skill(s) you intend to assess and the technique you will use to measure students' learning;

2. Develop a map which specifies the tasks and subtasks that are prerequisite to the assessed skill(s);

3. Write test items based on the identified blueprint and map, utilizing common conventions of item-writing;

4. Review the test items to confirm their match to the specifications and to assure that items do not contain extraneous complexities, unintended cues, or other technical flaws;

5. Field test the items to determine where item revisions are necessary, and/or where the blueprints and maps need to be adjusted; to determine whether there is a relationship between the hypothesized pre-requisites and the desired objectives; and to determine the number of items required for testing.

Each of these steps is described in the following sections.

Step One: Develop The Skill Blueprint

This first step of the test development process often is the most arduous. Developing a skill blueprint requires hard thought about the nature of the skill that is to be assessed and the nature of item content and format which can most appropriately assess its attainment. Because of the effort involved in test development and administration, these skills ought to reflect those which require large chunks of instructional time and which represent major goals for students for a unit, semester, or year.

Identify objectives worth testing. The first step within the specification process, then, is to identify objectives worth testing. A number of screens may be considered in determining the most suitable targets of assessment:

> How much instructional time does it take to teach the objective? As mentioned above, you'll want to select objectives that cover a reasonable amount of instructional time.
>
> How does the objective relate to other higher-order skills? Recent reports on the status of American education have been critical of the level at which some instruction occurs. Be sure that the knowledge and skills you are testing and teaching reflect or are pre-requisites to higher-level thinking, problem-solving skills and important educational goals.

How does the objective relate to long-term curricular goals? Like the concerns raised above, be sure that the objectives identified for assessment are relevant to important curricular objectives and are part of a coherent strand of learning.

What is the intrinsic importance of the objective? Related to all of the above, be sure that the objects of assessment reflect important and not trivial learning tasks.

Specify the skill required to meet the objective. After the objective(s) has been identified, try to clarify the nature of the skill(s) that students are expected to acquire. What are they supposed to be able to do?, e.g., comprehend the main idea of particular types of texts; solve particular types of physics problems; analyze the causes of particular types of world events; analyze particular literary works with regard to their plot, characterization, and setting; predict the short and long term consequences of particular environmental intrusions; recall major events of the civil war; write an expository essay with certain characteristics, etc.

Consider the level of cognitive complexity at which students are expected to function. For example, following Bloom (1956), does the skill of interest involve recall, application, analysis, or synthesis? Or, following Gagne (1970), does the skill represent concept learning (concrete or abstract), principles, procedures, or problem-solving?

Clarify the content which will be covered. Consider also the nature of the content that needs to be included on the test. You may want to examine available curriculum materials as well as your own judgment and experience as you consider some of the following questions:

In how many different contexts will students need to apply the skill? For example, in the reading example above, will students need to use their comprehension skill with expository and narrative texts, in texts where the main idea is implicit or explicit? In the physics example above, will students need to apply specific physics principles in laboratory settings, in real life-like situations in space, in aircraft, or in home situations? In the history example, how many and what types of historical events will students be required to analyze?

What information will students need to know? Is there a list of concepts, vocabulary, and facts, that students will be expected to acquire? For example, in the science example above, what bones will be included in the instructional and test content and what is their function? Or in the civil war example, what types of events are to be considered major?

How many different topics will be used to test students' skill acquisition? For example, in the expository essay example above, what kinds of prompts will be provided to students? Will they include topics which students have directly experienced, topics which are related to those learned in other parts of the curriculum, and will they require persuasion and/or description?

Are there pre-requisite skills that students will need to acquire?

The idea, again, is to clarify the nature of the skill or content that is to be assessed and diagnosed.

Select appropriate item type. Once you feel satisfied that you thoroughly understand the skill, consider what item format might be most suitable for the assessment and, within the selected format, what types of items are most appropriate.

Consider the range of item formats: selected response items, including true-false, matching, and multiple choice; constructed response, including short answer and essay; and performance measures, including observation and rating scales. There are no hard and fast rules for choosing particular item formats, although there is sometimes an inverse relationship between the ease with which an item is constructed and/or scored and its measurement validity. For example, although they're easy to construct, students have a 50% chance of guessing the correct answer to a true-false item. On the other end of the spectrum, although they are quite time consuming to score, essay tests provide the best measure of student's writing skill and are the only valid alternative where divergent responses are desired.

With regard to particular types of items within an item format,

brainstorm some alternatives and choose the one which can best elicit the skills in which you're interested. Write a couple of items which illustrate the kind of item you have in mind. If the items are to be administered via computer, be sure the item is structured to fit within the existing constraints, e.g., the number of lines that will fit in a single screen.

Write the skill blueprint. Once the sample items have been formulated, the skill blueprint can be written. Different researchers have suggested slightly different formats. The one described below combines models suggested by Popham (1980) and Baker (1974) and includes the following components:

- o General description - a brief description of the objective, skill, or knowledge to be measured.
- o Sample item - a model of what test items are to look like, including directions to be given to students.
- o Content limits - a description of the nature of the question that is to be presented to students.
- o Response limits - for selected response items, a description of the response options provided to students or for constructed response items, the set of rules or criteria that are to be used to judge the quality of a students' response.

The first two components are relatively straightforward: they include a statement of the objective selected for testing and instruction and the sample item that has been devised for assessing it. Include here also the directions that will be given to students. Explicit attention to the directions early on helps to assure that they will be clear and that students will understand how to complete each item.

Content limits describe the range of eligible content from which test items may be written. They may include rules for creating questions, and rules for the inclusion of prompts, cues, or additional materials such as

pictures, graphs, and reading selections.

Content limits for selected response items define and restrict the characteristics, format, and eligible content to be included in the item stem. By systematically including the different situations and contexts in which the skills are to be applied and/or the rules which define the assessed skill, test items can provide valuable diagnostic information, such as, in what situations are students able to demonstrate a particular skill?, what rules have students mastered? For instance, for a multiple choice item assessing students' skill in using appropriate pronouns, the content limits might be formulated as follows:

- The item stem will present the student with a short (3-5 sentence) paragraph which describes an action or event involving two or more named individuals.
- A blank will replace the named individual(s) in one sentence.
- Students will be asked to identify the pronoun which correctly completes the sentence.
- Items will be written to exemplify the following rules:

  When the pronoun is the subject of a sentence or clause, it should be in the nominative case.

  When the pronoun is the direct object, it should be in the objective case.

  When the pronoun is the indirect object, it should be in the objective case.

(Note that systematically including items reflecting each rule enables a test to diagnose which rule(s) is causing students' difficulty; the problem of ascertaining the number of items to be written to reflect each rule is addressed in a later section.)

Content limits for constructed responses define and restrict the prompt, the mode of response, and where appropriate, the conditions, setting or context surrounding the testing. The content limits for an expository essay task, for example, would specify rules for generating

essay prompts and the directions and any special cues to be given to students. For instance,

- The prompt will present students with a proposition and ask the student to take a position.
- The topic presented in the proposition to students must be one with which almost all high school students would be familiar, e.g., a topic dealing with a situation commonly encountered in daily living at home or at school.
- The topic must embody an issue on which students would be likely to have differing opinions, i.e., in favor or opposed to the proposition proposed.
- One sentence will provide brief background to the proposition and will include common reasons supporting the proposition. A second sentence will include common support for the opposing position.
- These sentences will be labeled: "Background."
- The backround sentence will be followed by the assignment which consists of the following sentence: "Write a paragraph in which you are in favor of, or opposed to, _____." Be sure to support the position you have taken.

Response limits provide rules for generating the correct response and incorrect alternatives for selected response questions and rules and criteria for judging the quality or correctness of a student's constructed response. Like the content limits, response limits help define the range of eligible content but here the focus is on student responses: what discriminations are expected and reasonable?; what are the characteristics of an acceptable response? what are common misconceptions? For selected response items, response limits provide rules for constructing the correct answer and the distractors, or wrong answer alternatives for each item. These rules should assure distractors that represent common student errors and which thus may provide important diagnostic information. For example, response limits for the pronoun example described above might be as follows:

- o Five alternatives will be provided for each item, the correct answer and four alternatives.
- o The correct response will exemplify the proper application of the given rules and will reflect the appropriate gender and number.
- o Distractors will consist of the following:

  - a pronoun in the correct case, but incorrect in number or gender;
  - a pronoun in the incorrect case, but correct in number and gender;
  - a pronoun representing an incorrect referent, but correct in case, number, and gender;
  - a pronoun in the incorrect case, incorrect in number and/or gender.

With such a set of alternatives, a student's wrong answer choice might provide information on whether he/she was having difficulty in identifying referents, was confused about case rules, and/or was having difficulties associated with number and gender.

For a constructed response item, response limits provide rules for judging or rating the adequacy of a student's response. Defining response limits using a set of concrete criteria maximizes both the diagnostic value of the assessment and its implications for instruction. For example, response limits for the writing example described above might be as follows:

> Student essays will be rated based on their organization, support, and mechanics. A five point scale will be used for rating each area, with a five designating the high end.
>
> Organization will be rated as follows:
>
> 5= essay is on topic; the paragraph includes a topic sentence which states a position regarding the assigned topic; the essay includes at least three reasons supporting the position; all sentences in the essay support the topic sentence.
>
> 4= .....
>
> Support will be rated as follows:
>
> 5= .....

The result of this specification process is a map for developing test items and likewise a map guiding instruction. Not only does the specification provide rules for developing multiple parallel test items, it likewise can be used to plan instuction and to generate relevant exercises for classroom practice, practice which will help students to acquire the specific skills they are intended to learn. So, although the process takes some time and effort, there are potential pay-offs.

Step Two: Specify The Skill Map

During the first step, the skill which is the target of assessment has been identified and well specified and a blueprint has been created for developing test items to assess that skill. The specification and the test items it implies, where possible, have been designed to provide diagnostic information about students' performance. For example, in the pronoun example cited above, the test items are to be created to assess students' attainment of particular rules of pronoun usage and the alternatives have been developed to provide information about whether students are experiencing difficulty with referents, case, and number. Likewise, in the essay example, scoring rules were created to rate students' writing in terms of concrete skills of organization, support, and grammar.

A finer grained diagnosis can be achieved by analyzing the level of difficulty at which students are able to operate, and/or the subtasks and subskills which they have mastered enroute to the desired assessed skill. In other words, suppose students are not able to correctly perform the assessed skill, is it possible to place them on a continuum from no skill through some skill to fully skilled, and how might one define the points on the continuum? If one can define the points on the continuum in terms of specific competencies and/or identify the relevant skill hierarchy, then it

is possible to devise test items which can appropriately diagnose students' skill level.

How to define the skill continuum or hierarchy is a problem that has been addressed by a number of researchers, but there are no hard and fast rules or breakdowns that are applicable across a range of subject areas. Combining logic, theory, and research in learning and instruction, as well as practical experience in teaching students the target skill, two inter-related strategies may be useful in diagnosing skill level:

- identify simpler contexts in which students' may be able to demonstrate the skill and/or simpler tasks which require skills similar to the target skill;
- identify pre-requisite skills and knowledge which students would need to master in order to attain the target skill.

Identifying simpler contexts/tasks. Several research-based principles can aid in the identification of simpler contexts or tasks which can help to define interim points on a skill continuum. These principles are inter-related rather than exclusive and include linguistic complexity, cognitive complexity, and level of discrimination.

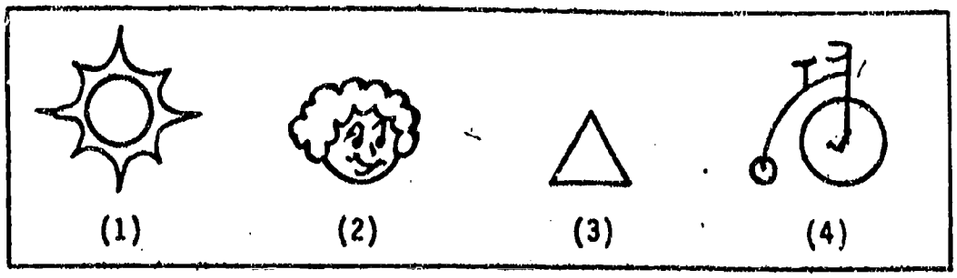
The logic is obvious of using linguistic complexity to help diagnose students' skills in reading. The question is, for example, if a student cannot comprehend the main idea of a particular passage, can he/she comprehend the same or different passage written at a lower level of linguistic complexity? Similarly, in an English example, if a student has difficulty analyzing the protagonist's character in a given story, can he/she perform the analysis with a simpler text? (It should be noted that when reading skill is not the object of assessment, linguistic complexity should be controlled, to the extent possible, so that it does not influence a student's performance; e.g., if a student's math skills were the subject

of assessment, the test developer would want to keep the language to a simple level so that reading ability did not influence performance.)

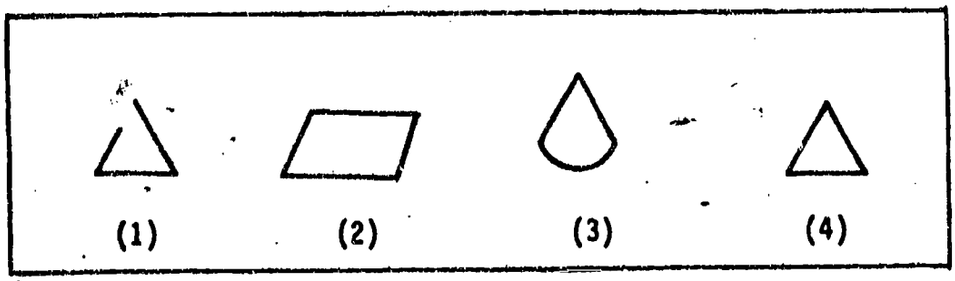
Cognitive complexity is a second factor which can help to define a continuum of task difficulty. It has to do with the level of processing required by a problem and the number of cognitive steps that a student would need to complete. The question is, if a student cannot handle one level of cognitive complexity, can he/she handle the problem at a simpler level of complexity? Consider, for example, the pronoun usage items described above. Students were to be given a short passage and were to be asked to identify the pronoun which would correctly complete a blank within the passage. The task requires students to use the context of the passage to identify the correct referent and then to match the referent with the appropriate pronoun. In a simpler task, the student might be given a sentence in which a subject or object were underlined and asked to identify the pronoun which could be substituted for the underlined word. Thus, the task would not require students to process the passage to identify the referent. Based on students' responses to these tasks, a teacher might be able to pinpoint a student's problem as related to identifying referents in context.

Required level of discrimination is a third factor which may be helpful in thinking about difficulty. Some tasks require fine levels of discrimination among concepts and topics while in other task only gross discrimination are necessary. For example, consider the following two items which ask students to identify a triangle from a set of alternatives (from Baker and Herman, 1983):

A.



B.



Both examples assess students' understanding of the concept "triangle," but the second clearly requires finer discriminations. To mark the correct answer in the second example, students must be able to apply features of three-sidedness, closed figure, linear figure, while in the first example knowing that a triangle is a geometric figure is sufficient information to arrive at a correct answer and any one of the three defining features of a triangle may be used to respond correctly.

Two other examples illustrate the notion of discrimination. Consider the question, "Which country is more democratic, Italy or France?" vs. "Which country is more democratic, the United States or the USSR?" Imagine also two literary analysis problems which ask students to describe the theme of given works. In one work, there is a unitary theme which is obvious; in the second, there are several sub-themes and the central one is less salient.

Closely related to the level of required discrimination is the level of prompting, and/or salient cues given to the student about what he/she is supposed to do. Suppose, for example, that a student is given life-like problems which he/she is supposed to solve using principles of physics, but the problems are silent on which principle(s) apply. A simpler version might prompt the student on what principle to use for each problem.

Identifying Prerequisite Skills and Knowledge. The previous section

has discussed potential strategies for simplifying the context in which students apply their skill in order to diagnose the level at which they are able to operate. A complementary (and inter-related) approach to the diagnostic problem is to consider simplifying the skills which students are asked to demonstrate and to try to locate their performance within a hierarchy of prerequisites identifying gaps where instruction is needed. For example, in the reading comprehension example cited above, students were asked to identify the main idea of given passages. Skills prerequisite to comprehending the main idea and which might be assessed are comprehending details of the passage, understanding the specific vocabulary used in the passage, and so on.

The iterative question which needs to be addressed in identifying prerequisite skills is "What does a student need to be able to do in order to attain a given skill? What subskills does he/she most likely need to learn enroute to the desired skill?" Examples easily come to mind in the area of mathematics, (cf. Gagne, 1977). In order to subtract whole numbers of any size, a student would need to be able to, in ascending order of difficulty, to subtract without borrowing, to subtract when several borrowings are required in non-adjacent columns, and when successive borrowings are required from adjacent columns.

In order to apply task analyses in other areas of the curriculum, think about the nature of the skill you are assessing. What rules, procedures, and/or principles does a student need to know in order to attain the skill? What concepts does he/she need to understand? Are there particular facts that need to be accessible? Each of these represent potential diagnostic points on the skill hierarchy.

The skill map. Use the above strategies, combined with practical knowledge about likely and/or common sources of students' problems and

errors to develop a map, or continuum, which can guide diagnostic testing. How many subtasks or subskills should be included in the map? On the one hand, the more that are included the greater the diagnostic potential of the test. On the other hand, each diagnostic point adds greatly to test development and administration time and it may not be feasible to use more finegrained information in instruction. For example, it may be faster and logistically easier to carefully reteach the skill to a group than to painstakingly uncover the unique problems of each student. The number of diagnostic points included on the test, then, will relate to both feasibility and potential utility. Probably these two conditions will suggest that a couple of points reflecting common levels of student performance are reasonable for assessment.

Once the diagnostic assessment points have been specified on the skill map, then the types of items that will measure each point need to be identified. Ideally, this process mirrors the process described above for developing a skill blueprint, with blueprints being devised for each subskill and/or subtask. Time constraints, however, may limit the level of detail included.

Step Three: Develop Test Items

Once the skill blueprints have been specified, developing the test items is a matter of simply following the specified rules. How many items need to be created? The statistical analysis conducted late during step five will provide a good estimate of the number of items that will need to be included on the final version of the test. At this preliminary stage, however, the answer is "as many as possible," and at least three to five items for each diagnostic point on the test, i.e., 3-5 items for each subskill and for each rule and/or task context included within the skill blueprints.

In addition to following the test specifications, item writers will also want to keep in mind conventional rules of thumb which help prevent the inclusion of extraneous factors that can confound an examinee's response. These rules concentrate on factors such as linguistic, semantic, and grammatical features that may enable an unknowing student to give a correct response or that may prevent a knowing student from responding correctly. Gronlund (1968) and Conoley and O'Neil (1979) provide a thorough explication of such rules. Typical rules are summarized in Figure 1.

Step Four: Review the Test Items Once items are developed, the next step is to conduct a thorough review, considering two basic questions:

- Do the items match their specification?
- Are they free from technical flaws, i.e., do they follow conventional rules of item construction?

Do the items match their specification? The answer to this question is critical to establish the content validity of the test. The process is straightforward: have each item examined by a colleague to compare its match with each element in the specification. That is, the description of eligible subject matter and item features provided in the content limits needs to be compared with the content and features of the test question; and the specification rules for creating correct and incorrect answer alternatives must be compared with the actual set provided in selected response items. The items should be checked also to see that they follow the prescribed format and that appropriate directions are given. While covered again under "technical flaws," check also to assure that the language used in the items is not unnecessarily difficult or complex and that items are free from content that might be biased against particular groups of students. Where any problems are encountered, suitable item

Figure 1
General Guidelines for Item Writing

Typical Rules for Multiple Choice Items:

1. The stem of the items should be meaningful by itself and should present a clear problem.
2. The stem should be free from irrelevant material.
3. The stem should include as much of the item as possible except where an inclusion would clue the responses. Repetitive phrases should be included in the stem rather than being restated in each alternative.
4. All alternatives should be grammatically consistent with the item stem and of similar length, so as not to provide a clue to the answer.
5. An item should include only one correct or clearly best answer.
6. Items used to measure understanding should contain some novelty and not merely repeat verbatim materials or problems presented in instruction.
7. All distractors should be plausible and related to the body of knowledge and learning experiences measured.
8. Verbal associations between the stem and correct answer or stereotyped phrases should be avoided.
9. The correct answer should appear in each of the alternative positions with approximately equal frequency and in random order.
10. Special alternatives such as "none," "all of the above" should be used sparingly.
11. Avoid items that contain inclusive terms (e.g., "never," "always," "all") in the wrong answer.
12. Negatively stated item stems should be used sparingly.
13. Avoid alternatives that are oppositive in meaning or that are paraphrases of each other.
14. Avoid items which ask for opinions.
15. Avoid items that contain irrelevant sources of difficulty, such as vocabulary, sentence structure.
16. Avoid interlocking items, items whose answers clue responses to subsequent items.
17. Don't use multiple choice items where other item formats are more appropriate.

Figure 1 (continued)

Typical Rules for Short Answer and Completion Items:

1. A direct question is generally better than an incomplete statement.
2. Word the item so that the required answer is both brief and unambiguous.
3. Where an answer is to be expressed in numerical units, indicate the type of units wanted.
4. Blanks for answers should be equal in length. Scoring is facilitated if the blanks are provided in a column to the right of the question.
5. No grammatical cues should be give, e.g. a _____; an _____.
6. Where completion items are used, do not leave too many blanks.
7. For completion items, only key words should be left blank. Leave blank only those things that are important to remember.
8. In composing items, don't take statements verbatim from students' textbook or instruction.
9. The scoring key should anticipate possible synonyms or acceptable variants at the desired response.

Typical Rules for True-False or Alternative Response Items:

1. Avoid broad general statements for true-false items.
2. Avoid trivial statements.
3. Avoid negative statements and especially double negatives.
4. Avoid long complex sentences.
5. Avoid including two ideas in a single statement unless cause-effect relationships are being measured.
6. Avoid questions which include indefinite terms, degrees or amounts.
7. Include opinion statements only if they are attibuted to particular sources.
8. True statements and false statements should be approximately the same length.
9. The number of true statements and of false statements should be approximately equal.
10. Avoid taking statements verbatim from students' text or instruction.
11. An item's truth or falsity should not depend on an insignificant word or phrase.

revisions need to be made, e.g., changing language, reducing ambiguity, changing the item stem or alternatives to match the blueprint. (In some cases where unanticipated problems emerge, there may be instances where the blueprint needs to be changed and/or modified.)

Are the items free from technical flaws? The review process here is also straightforward. Simply check the items against the general rules for constructing test items of particular types, and where flaws are detected; correct them. As with the content review described above, it is preferable to have the review conducted by a colleague, yielding the advantage of having a "cold," objective eye.

Step Five: Field Test the Items

Field testing the items is a final step in the test development process to assure high quality items, to verify the test structure, and to determine the number of items that will be needed to reliably diagnose students' performance. The optimal field test procedures involve a two stage process: 1) pilot test the items with a small sample of students to check their appropriateness; 2) administer the test to a larger sample to validate the subskills that need to be included in the test and the number of items required for each skill and subskill.

The initial pilot test. The purpose of the first pilot test is to determine whether the items are appropriate for students and to identify items that are potentially in need of revision. Have a small number of students who are similar to the intended student population take the entire test and provide feedback on any problems they encounter, e.g., vocabulary or directions that are unclear; items where there seem to be more than one (or no) right answers. This feedback helps indicate where revisions are necessary.

Item difficulty indices (the percent of students who answer an item

correctly) also help signal potential problem items. Because they are based on the same blueprint, one would expect similar item difficulties for all items measuring the same subskill or task. Gross deviations indicate items which need additional review. For example, suppose that item difficulties for four of the five items measuring a particular subskill are .5-.7; however, the difficulty of the fifth item is .25. This latter item should be re-examined to determine whether it is aberrant and is unintentionally confusing the correct response, whether it matches the specification, whether it represents a problem type that is different from the oth items, whether the correct answer has been miskeyed, and/or whether there are typographical or other errors in the items. Any detected errors or deviations will need to be corrected. Item difficulties can also be used to help judge the appropriateness of the test for particular students. In order to be useful in a diagnostic sense, a test should measure target skills which are difficult for a substantial number of students: If all or most students get all or most of the items correct, there is little to diagnose.

The field test. Once the initial piloting has been completed and revisions made, the revised version of the test needs to be field-tested by administering it to a larger sample of students (at least 100 per student-population). Student performance on this field-test should then be analyzed to establish the technical characteristics of the test and to direct further the revision process. While a thorough description of appropriate analytic procedures is outside the scope of this report, the use of generalizing analyses is recommended for the field test analysis. Although such analyses are complex and will require the services of an

expert statistician, they provide full information on the structure and reliability of the test. Two types of generalizability analyses, identified below, are recommended. A brief rationale for these analyses follow:

- o generalizability analyses to analyze the structure of the test and to determine which subskills and tasks have distinct diagnostic value and to verify hierarchical relationships among skills and sub-skills;
- o separate generalizability analyses for each skill and task in the profile to determine the number of items that should be included in the test to obtain a reliable measure of those skills.

Generalizability analyses related to structure. The content and/or skill dimensions included on the test reflect hypotheses about what causes students' performance to vary in a particular skill area, and about why some students score very highly and others do not. These hypotheses are validated if one can demonstrate that student performance within a dimension is relatively consistent and reflects a uniform (sub)skill, but is inconsistent, or varies, across dimensions. Under these conditions, a particular student's total score is "explained" by his/her subskills, e.g., a student performs at a certain level because he/she scores consistently well in some (or all) subskills and is consistently unable to perform others. These latter skills represent those in need of remediation.

A content or skill dimension (including rules and contexts within the skill blueprint and tasks and subskills included in the skill map) has diagnostic utility and needs to be represented on a test if it demonstrates such explanatory power. In the absence of such power, knowledge about student performance on the dimension provides little additional information to teachers. That is, if students' performance is inconsistent within an

area, then this area does not represent a single distinct skill. Or, if students perform at the same level on all dimensions, then there is no need to profile them separately or to provide separate scores.

Generalizability analyses can be used to determine which of the dimensions included on the test have explanatory power and therefore should be retained as separate subskills. The analysis treats each dimension as a separate factor and examines the amount of variance it contributes to the total score. While there is no rule of thumb about what proportion of variance represents a large amount, some researchers have recommended 3.5-5% as a cut-off. The decision involves a trade-off between cost and information. Using a small proportion as a minimum may produce more detailed skill profiles than are necessary. Using a large proportion as a minimum, on the other hand, may cause important sources of student problems to be overlooked or disregarded.

Generalizability analyses related to number of items.

Generalizability analyses can also be used to determine the optimal number of items to include for each content or skill dimension covered on the test. The analytic question is "how many items are needed to provide a generalizable or reliable measure of student performance?" and separate analyses are conducted for each content or skill dimension. Like the analyses above, there is no firm rule of thumb for how reliable or consistent a score needs to be, although coefficients of .6-.7 are common. (See Webb et al, 1983 for a fuller explantion of the use of generalizability analysis.)

Based on these analyses, the final diagnostic test can be constructed, reflecting the structure and item requirements indicated by the above analyses.

Summary

Diagnostic testing can provide specific information about student skills as a decision-making aid to teachers in prescribing instruction, identifying needs for remediation, determining effective instructional materials and methods, and ultimately, improving student learning. Diagnostic testing, as viewed here, includes individual and group assessment of students' skills in specified cognitive domains. A methodology is presented for designing diagnostic tests which assess the extent of student learning and are sensitive to sources of difficulty within a skill or context area. This 5-step methodology for diagnostic test development includes:

1) Developing a skill blueprint including a general description of the objective or skill, a sample item, content limits, and response limits;

2) Specifying the skill map including sub-skills or simpler contexts which students should masten enroute to the desired skill under assessment;

3) Formulating test items that match specifications and follow conventions for sound item-writing;

4) Reviewing test items to insure match to specifications and technical quality;

5) Field testing the items and revising to insure that the test is appropriate for the intended student population and structured to provide meaningful and reliable diagnostic information.

Bibliography

Baker, E.L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. In Hively, W., Domain-referenced testing. Englewood Cliffs, New Jersey: Educational Technology Publications, 1974.

Baker, E.L & Herman, J.L. Task structure design: Beyond linkage. Journal of Educational Measurement, V. 20, N. 2, Summer 1983

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

Conoley, J. & O'Neil, H. A prime for developing test items. In O'Neil, H. (Ed.), Procedures for instructional systems development. New York: Academic Press, 1979.

Gagne, R.M. Analyses of lectures. In L. Briggs (Ed.) Instructional design: Principles and applications. Englewood Cliffs, New Jersey: Educational Technical Publications, 1977.

Gagne, R.M. The conditions of learning. (2nd edition) New York: Holt, Rinehart, and Winston, 1970.

Gronlund, N.E. Constructing achievement tests. Englewood Cliffs, New Jersey: Prentice Hall, 1968.

Popham, W.J. Domain specification strategies. In Berk, R.A. (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: Johns Hopkins University Press, 1980.

Webb, N.M., Herman, J., & Cabello, B. Item Structures for Diagnostic Testing. Center for the Study of Evaluation, UCLA Graduate School of Education.