DOCUMENT RESUME

ED 121 280 IR 003 257

AUTHOR Geisinger, Kurt F.
TITLE A Systems Approach to Item Production and Review in a Computer Managed Instruction Project.
PUB DATE Apr 76
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS *Computer Assisted Instruction; *Criterion Referenced Tests; Higher Education; Inservice Education; *Item Banks; Nursing; Program Descriptions; *Test Construction; University Extension
IDENTIFIERS Computer Managed Instruction; Computer Managed Review and Examination; *Item Generation

ABSTRACT

An item generation procedure is described which was utilized in the development of Computer Managed Review and Examination courses for the education of nurses in remote areas. The major emphases are the processes of domain definition, item writing, and item edition. Specific discussion is presented concerning methods of item construction to assess technical vocabulary, concept learning, and the application of nursing principles to the solution of problems. The entire test construction procedure is briefly reviewed; this procedure includes numerous quality checks to insure the production of both high caliber instructional materials and domain-referenced tests. The criteria used at various editing and review stages are mentioned. An initial evaluation of the items is made, and problems inherent in the item generation procedure are offered. (Author/JY)

A SYSTEMS APPROACH TO ITEM PRODUCTION AND REVIEW IN A COMPUTER MANAGED INSTRUCTION PROJECT*

Kurt F. Geisinger
The Pennsylvania State University

Abstract

This paper describes the item generation procedure utilized in the development of Computer Managed Review and Examination courses for the education of nurses in remote areas. The major emphases are the processes of domain definition, item writing, and item editing. Specific discussion is presented concerning methods of item construction to assess technical vocabulary, concept learning, and the application of nursing principles to the solution of problems. The entire test construction procedure is viewed; this procedure includes numerous quality checks to insure the production of both high calibre instructional materials and domain-referenced tests. The criteria used at the various editing and review stages are mentioned. An initial evaluation of the items is made and problems inherent in the item generation procedure are offered.

Since the mid to late 1960's, traditional achievement testing has been the subject of considerable criticism and innovation. Glaser (1963), Bormuth (1970), and other measurement experts have strongly encouraged the educational community to re-evaluate the testing procedures used in instructional programs. The construction and selection of achievement test items, in particular, has been a focus of attention.

The problematic nature of achievement test and item construction rises to even greater prominence in Computer Managed Instruction (CMI) and other individualized instruction programs (Anderson et al.: 1974,

*Presented at the 1976 AERA Annual Meeting, San Francisco, California, April, 1976.

Mitzel, 1974). As Robert Gagne (1975, p. 145) has said, "To the extent that modern educational trends, at all levels, favor self-education and independent learning, the means of observing and assessing the outcomes of learning becomes a matter of considerable importance."

The importance of item construction is exemplified in the fact that this CMRE project involves the development of eight nursing courses, each of which necessitates the construction of approximately 1,100 items[1], for a project total of almost 9,000 items!

Typically, testing in education serves a single prime purpose: to accredit or certify competence. Within the CMRE model of this project, however, four distinct purposes can be delineated for testing. First, the initial test serves a placement function. Then, the review questions throughout the instructional program serve two functions: to diagnose student learning and prescribe remedial instructional materials (and hence, to keep students from taking tests before they are adequately prepared for them) and to maintain student interest, motivation, and attentiveness. After instruction has been completed, the final examination serves the traditional credit awarding function.

Achievement measures are evaluated in terms of content validity. Most authors in the field of measurement (Cronbach, 1971) recognize that content validity is assessed largely with respect to the degree

[1]Of these 1,100 items, approximately 700 are test items and 400 are for review. Although the differences between these items is often not large, they are kept as two separate item pools. Since review items follow instruction more closely chronologically, we have tended to use these items with more specificity.

to which logical and systematic test construction procedures are utilized. The prime goal of this paper is to state the test construction procedures employed explicitly. Additionally, many individuals (Baker, 1974; Ferguson, 1972; Glaser, 1970; Hambleton, 1974; Nitko, 1974; Willingham and Geisinger, 1976) have been interested in the degree of parellelism between educational measurement practices and the other components of instructional systems. Only through such explicit statements of procedures can procedural inconsistencies ("working at cross-purposes") be identified and removed.

Developing the CMRE Courses

Staffing

Individual faculty members from the College of Human Development's Department of Nursing were assigned as course authors for each of the eight CMRE courses. Additionally, a nurse-research assistant was hired one-half time to assist each author. While these assistants aided the authors in reviewing academic materials for use "off-line," their major role was that of item writing. A professor from the Department of Educational Psychology conducted several item writing workshops for these individuals. Two graduate students in Educational Psychology served as principal item reviewers or editors, and wrote items upon occasion.

The Course Development Process

The following eleven step course development sequence was utilized in the preparation of all eight courses to facilitate progress and to

keep track of the location of items within the organization.

1. Instructional Material Developed
2. Initial Test Item Construction
3. Initial Item Review
4. First Item Revisions Made
5. Item Typing on Paper
6. Second Item Review
7. Second Item Revisions Made
8. Magnetic Tape Selectric Typing
9. Course Author Item Approval Granted
10. Magnetic Tape Corrected
11. On-line Review and Revision

Instructional Material Developed

The development of instructional materials for these courses is primarily a two-step process. Each author first enumerates his major goals for the course. From this list of goals, a detailed subject-matter outline is constructed. Using this outline, a committee of the Department of Nursing then judges the adequacy of course coverage. The second step involves the operationalization of the outline; appropriate materials (texts, articles, films, tapes, etc.) are developed or selected to represent the topic areas listed in the outline. (This procedure is crucial for the test construction process; Appendix A describes how the instructional material selection and development relates to current measurement topics such as universe and domain definition, domain-referenced testing, and criterion-referenced testing.) Then, this body of curricular information is divided into single study session-sized segments called lessons. Each lesson is weighted according to its importance, this weighted importance being directly proportional to the eventual number of items for that lesson.

The Use of Summary Statements

Consistent with the procedures used at the University of Illinois'

CAICMS project, subject matter experts devise summary statements from the instructional material (Wietecha and Anderson, 1974). Summary statements are abstractions of the major themes within instruction and of those elements of critical importance within each lesson. A single lesson might have as many as twenty to twenty-five such statements. These summary statements are written verbatim from the textual material on which the lesson is based. The specific subject and predicate are unchanged from the exact wording of the text; the summary statements are kept as consistent with the language of the text as possible. The length of the summary statements ranges from a single sentence to a short paragraph.[1] Each summary statement is referenced by module, lesson, and page of text or article within the lesson.

Writing the Test Items

After attending the intensive objective item-writing workshops, the subject matter experts involved in each course construct the bulk of the questions for that particular course. (By subject matter experts, both course authors and course assistants are indicated.) Each item is constructed from a summary statement. Like each summary statement, each item is referenced to the page of the written instruction from whence it came. This is used later in the formation of diagnostic-prescriptive statements for the examinees.

[1] In the interest of increasing efficiency, after the first year of the project, some course authors have dispensed with the use of summary statements. In their place, the authors "highlight" or underline those statements in the instructional materials themselves. This, of course, saves the time of copying the statement verbatim from the book. Furthermore, it permits the author to select those topics on which he desires items, while allowing the course assistant to actually construct the items from the underlined passages. This, too, represents a time-saving.

Basically, four distinct types of items typify all the items used on the tests. The first is used if the summary statement includes a specific term or name which the subject matter expert believes the student should be able to recall (rather than recognize.) A constructed response item, in which the student is required to type in the term, is used. In this case, however, the stem of the item is altered from the verbatim summary statement through the use of paraphrasing. Sometimes, grammatical transformations are also performed on the paraphrased summary statement. This insures that the item is semantically encoded in memory, not merely orthographically or phonologically encoded (see Anderson, 1972, for an in depth explanation of this point.) Item writers must be reminded often that this type of item is only appropriate for specific words or phrases. It is a relatively easy kind of item to construct because there is no need for distractors; it is simply not justified for testing those general concepts with many synonyms. Generally, no more than two or three synonyms are allowed to be keyed as correct. Approximately ten per cent of our items are of this variety.

A second type of item assesses the ability to employ nursing principles. Generally, such a principle recommends a course of action to the nursing student which is appropriate under certain circumstances. The stem of such a question represents one such specific circumstance. The nurse is asked what to do. The nurse must correctly apply the principle to this new situation. Unless, as is infrequently the case, the desired response is embodied by a specific term, a multiple-choice format is used, with a number of possible actions listed as options.

Preferably, each such option represents a different orientation or principle. All options are mutually exclusive. The nurse must select which action is best. This type of item may also be used for computational problems. (For example, calculations of proper dosage are important in some nursing courses.) Most often, the examinee must type in the correct response in such computations. This controls for the possibility that the examinee could "work back" from the options to discover the correct answer. Any problem, either verbal or numerical, presented to the nurse is new, different from any examples given as part of instruction. This insures that the student must determine the answer by applying the principle, rather than answering from rote. The Illinois CAICMS project referred to these principle-testing items as application items (Wietecha and Anderson, 1975). Both titles seem equally appropriate. Principle-testing questions account for approximately fifteen percent of the CMRE nursing questions.

The third variety of test items tests the student's mastery of a concept. Most often, the student is presented with a number of examples. Here, the student must choose which of the examples are instances (examples) of the concept. These items always include both positive and negative instances, thus forcing the student to perform a discrimination in the demonstration of his mastery. Options are generally not mentioned in the text, but newly constructed. This helps insure that the student has learned the concept, not just memorized those instances as used in the instructional material. About twenty-five to thirty percent of the items are of this variety.

The final and most frequently occurring type of item does not specifically assess concepts, principles, or terms, as do the previous item types. Rather, this type of item is simply a paraphrase of the summary statement, with an element, usually the subject, deleted. The task of the student is to recognize a paraphrase of the deleted part among the options. In some cases, when the subject is important in its own right, but the subject matter expert does not feel the nurse must recall the term specifically, the words are listed verbatim as options. (An example of this would be the titles of each of the eight stages of development in Erik Erikson's theory.) Sometimes the predicate is tested rather than the verb; an effort is made to test the most important aspects of the summary statements.

Item Writing Rules

The project attempted to avoid absolute rules concerning item writing. Several such rules did emerge, however. True-false questions are not permitted. Few concepts or principles are ever purely true or false. That such items are correctly guessed quite frequently argued against inclusion of either true-false questions or multiple choice questions with only two or three options. Use of the options "none of the above," "all of the above," and combination responses (i.e., and c) are not allowed. Use of such words as "always," "only," and "never" in options are avoided, as are other "specific determiners" or extraneous clues (Davis and Diamond, 1974). Questions aimed at tricking the student, or forcing him to make overly fine discriminations, are discouraged.

Item Format

As mentioned above, most items used are either multiple choice or short answer constructed-response format. In general, the rule determining which of these is used concerns the necessity of the student's being able to recollect the specific term or answer. As mentioned previously, two separate pools of items are kept: test items and review items. There tends to be a larger proportion of constructed response review items than test items. Matching questions are not used as test questions because of scoring problems. However, since one of the goals of review questions is to maintain student interest and attentiveness and because students tend to enjoy such items, matching items are used in review sections.

One of the advantages of the computer system is the use of multiple choice items which have more than one option as correct. These are used primarily as items testing a student's understanding of a concept. Special instructions concerning the student's response accompany these items. A typical such item would be "Which of the following are symptoms of pneumonia? Select one or more correct answer." Clearly, a student is less likely to answer this type of question correctly by guessing.

Editing the Test Items

Two graduate students in educational psychology serve as item editors: one performs the first editing, the other, the second. In reading each item, one of the following four judgments are made, and then the item is returned to the author.

1. "The item is fine, acceptable as it is."
2. "The item has problem 'X', here is a revision. How does that seem to you?"
3. "The item has problem 'X'. I suggest you make the following changes.
4. "The item is poor for the following reason(s). I suggest you start over again with the summary statement. Write another item."

The prime job of these editors is to analyze the items according to accepted rules of good objective item writing (Davis and Diamond, 1974; Tinkelman, 1971; Wesman, 1971; Wood, 1961). The item editors also re-paraphrase items to make them more straightforward, clear, and less reliant on the vocabulary of the text or instructional material. Frequently, after such item revisions, an item iterates between author and editor several times before both individuals are satisfied that the item is of acceptable content and form. The item editor also attempts to analyze the examinee response called for and attempts to determine whether this is congruent with the purpose of the item. If, for example, the item writer requires examinees to select the name of an appropriate drug from a list of five, the goal of such an item may be better served by a question of the constructed response format. On the other hand, if the item writer wishes the examinee to respond with a general concept, especially if that concept is referred to by various synonyms, a multiple choice format is called for.

The retyping of items between the two editorial processes helps to keep the judgments independent. The use of two editors is primarily in the interest of quality control.

Before each item is placed "on-line," the course author makes a final item approval. This allows the author to view all the items of a module or lesson at a single time and to make more global formative recommendations. Once "on-line," a member of the project staff checks each item, insuring that the item has been correctly keyed and the programming operable. These checks prevent faulty material from being sent to a mobile instruction site.

Evaluation of the Test Construction Process

Face Validity

Face validity refers to how well the items appear to be measuring the subject matter. Test construction experts have tended to consider face validity only to the extent to which it is needed to sell a test. Face validity has heightened importance for CMRE questions. The reason for this is that the students' prime interaction with the instructional system is in answering the questions. If the students perceive the questions as being trivial or irrelevant, they will lose respect for the potential usefulness or importance of the instruction. For these reasons, the relatively high number of realistic problems included in the examinations for the student to solve appears an extremely favorable quality. Furthermore, because the system follows a diagnostic-prescriptive model, the student is not simply told he has failed; he is told in what aspects of a lesson he needs further study.

Content Validity

As mentioned earlier, content validity is largely assessed in viewing the test construction process systematically, and judging how adequately the test items represent the domain. The domain has been carefully defined and summary statements have been made extremely consistent with the instructional materials. Then, the statements are paraphrased and often grammatically altered such that a student must comprehend the instructional material to answer it correctly. Quality control is assured in that all items are read several times by several different people before the items go "on-line." A faculty committee of the Department has evaluated all course outlines as adequate representations of the subject matter. Furthermore, nationally known nursing experts are being brought to Penn State to evaluate the CMRE courses.

Problems

The test construction process appears to be largely successful. However, several problems do appear worthy of mention.

1. The fact that the item editors were largely unable to make nursing-related statements leads to inefficiency: this is especially troublesome in the attempt to generate plausible alternatives for the mulitple choice items.
2. Even with the utilization of carefully chosen off-line instructional material, academic idiosyncracies on the part of the text authors are found. In constructing questions to test

such material, the item writer is forced to preface the item with, "According to . . ." One of the goals of CMRE is to allow students who have previously learned material to bypass the instruction on it a second time. Material which is textbook - or author-specific is not conducive for this purpose.

3. When starting with summary statements, a writer finds that he can generate a considerable number of items from a single summary statement. Selecting which item is best is an extremely unscientific process. This is especially difficult when the different items appear to have widely different levels of difficulty.
4. The procedure of determining what an item assesses (a concept, a principle, a term, etc.) is a highly subjective, mentalistic process, subject to disagreement among item writers.

These are problems requiring practical solutions. As increasing numbers of CMI and CMRE projects are developed, we hope that such assessment problems can be solved or handled in a better manner. Of course, the ultimate beneficiaries of such solutions are not the future CMI developers, but the future students (Popham, 1974).

REFERENCES

1. Alkin, M. C. "Criterion-referenced measurement" and other such toms. In Harris, C. W., Alkin, M. C., and Popham, W. J. (Editors), Problems in criterion-referenced measurement, Center for the Study of Evaluation, University of California, Los Angeles, 1974.

2. Anderson, R. C. How to construct achievement tests to assess comprehension. Review of educational research, 1972, 42, 145-170.

3. Anderson, T. H., Anderson, R. C., Dalgaard, R. B. Wretecha, E. J., Biddle, W. B., Pader, D. W., Smock, H. R., Alessi, S. M., Surber, J. R., and Klemt, L. L. A computer based study management system. Educational Psychologist, 1974, 11, 36-45.

4. Baker, R. L. Measurement considerations in instructional product development. In Harris, C. W., Alkin, M. C., and Popham, W. J., (Editors), Problems in criterion-referenced measurement, Center for the Study of Evaluation, University of California, Los Angeles, 1974.

5. Bormuth, J. P. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

6. Cronbach, L. J. Test validation. In Thorndike, R. L. (Ed.), Educational measurement, (2nd ed.), Washington, D. C.: American Council on Education, 1971.

7. Davis, F. B. and Diamond, J. J. The preparation of criterion-referenced tests. In Harris, C. W., Alkin, M. C., and Popham, W. J., (Editors), Problems in criterion-referenced measurement. Center for the Study of Evaluation, University of California, Los Angeles, 1974.

8. Donlon, T. F. Some needs for clearer terminology in criterion-referenced testing. Presented at the annual meeting of the National Council for Measurement in Education, 1974.

9. Ebel, R. Some limitations of criterion-referenced measurement. Paper presented at the annual meeting of the American Educational Research Association, 1970.

10. Ferguson, R. L. A model for computer-assisted criterion-referenced measurement. Education, 1970, 81, 25-31.

11. Fremer, J. Development of school-based criterion-referenced testing systems. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D. C., April 2, 1975.

12. Gagne', R. M. Observing the effects of learning. Educational Psychologist, 1975, 11, 144-157.

13. Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

14. Glaser, R. Evaluation of instruction and changing educational models in Wittrock, M. C. and Wiley, D. C. (Editors). The evaluation of instruction: Issues and answers, New York: Holt, Rinehart, and Winston, Inc., 1970.

15. Glaser, R. and Nitko, A. J. Measurement in learning and instruction. In Thorndike, R. L., Educational Measurement, (2nd edition). Washington, D. C.: American Council on Education, 1971.

16. Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. Review of educational research, 1974, 44, 371-400.

17. Hambleton, R. K., Swaminathan, H., Algina, J. and Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. An invited symposium presented at the annual meeting of the American Educational Research Association, Washington, D. C., April, 1975.

18. Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 5-10.

19. Jackson, R. Developing criterion-referenced tests. TM Report No. 1, Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1970.

20. Klein, S. P. and Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. TM Report No. 26, ERIC Clearninghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, New Jersey, 1973.

21. Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

22. Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.) Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Co., 1974.

23. Mitzel, H. E. (Ed.) An examination at the short-range potential of computer-managed instruction. Conference Proceedings, November, 6-8, 1974.

24. Nitko, A. J. Problems in the development of criterion-referenced tests: The IPI Pittsburgh experience. In Harris, C. W., Alkin, M. C., and Popham, W. J. (Editors), Problems in criterion-referenced measurement. Center for the Study of Evaluation, University of California, Los Angeles, 1974.

25. Popham, W. J. Selecting objectives and generating test item for objectives-based tests. In Harris, C. W., Alkin, M. C., and Popham, W. I. (Editors), Problems in criterion-referenced measurement. Center for the Study of Evaluation, University of California, Los Angeles, 1974.

26. Popham, W. J. and Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

27. Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-147.

28. Skager, R. W. Generating criterion-referenced tests from objectives-based assessment systems: unsolved problems in test development assembly, and interpretation. In Harris, C. W., Alkin, M. C., and Popham (Editors), Problems in criterion-referenced measurement. Center for the Study of Evaluation, University of California, Los Angeles, 1974.

29. Tinkelman, S. N. Planning the objective test. In Thorndike, R. L., Educational Measurement, (2nd ed.) Washington, D. C.: American Council on Education, 1971.

30. Wesman, A. G. Writing the test item. In Thorndike, R. L., Educational Measurement, (2nd ed.) Washington, D. C.: American Council on Education, 1971.

31. Wietacha, E. J. and Anderson, R. C. The preparation of test items to maintain attentive study behavior. In Anderson, R. C. and Anderson, T. H., Development and Implementation of the CAI Study Management System (CAISMS). Preliminary technical report, March, 1975.

32. Willingham, W. W. and Geisinger, K. F. Developing an operational model for assessing experiential learning. In Willingham, W. W. and Nesbitt, H. S., Implementing a program for assessing experiential learning. Princeton, New Jersey: Cooperative Assessment of Experiential Learning, Educational Testing Service, 1976.

33. Wood, D. A. Test construction. Columbus, Ohio: Merrill, 1961.

APPENDIX A

INSTRUCTIONAL DEVELOPMENT FOR DOMAIN-REFERENCED TESTING

Instructional Development for Domain-Referenced Testing

Currently, considerable inconsistency in educational measurement vocabulary exists, especially with respect to criterion-referenced testing (Alkin, 1974; Donlon, 1974; Millman, 1973, 1974). These test construction experts have argued that the term, criterion-referenced testing, should refer only to those tests where items are referenced to either behavioral objectives or amplified behavioral objectives. On the other hand, a domain-referenced test is "any test consisting of a random or stratified sample of items selected from a well-defined set or class of tasks (a domain)" (Millman, 1974)[1]. On such a measure, each examinee is measured to discover the degree to which he has attained the intents of instruction and not to see how he compares with other examinees with respect to his capacity to learn the instructional material. Millman further argues that such domain-referenced tests yield scores which are unbiased estimates of the percentage of all items within a domain mastered, written or unwritten. Such scores are extremely desirable for both placement and crediting decisions, and for insuring content validity. With a well-defined domain, there should be high agreement among experts as to what constitutes membership within the domain; Shoemaker (1975) has argued that a "universe" of all knowledge within an academic discipline must become operationally defined as a domain. In this CMRE project, the boundaries of the item domain have been so operationalized, as the set of "off-line" instructional materials: texts, articles, pages, films, tapes, etc.)

[1]Whereas it is popular, currently, to have criterion-referenced tests in instructional projects, Ebel (1970, p. 5) has demonstrated that "in areas where the emphasis is on knowledge and understanding, the effective use of criterion-referenced measures seems less likely."