DOCUMENT RESUME

ED 250 387 TM 840 738

AUTHOR Hanson, Ralph A.; And Others
TITLE Development and Verification of Instructionally Sensitive Achievement Tests.
INSTITUTION Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO SWRL-TR-69
PUB DATE 15 Dec 80
CONTRACT NEC-00-3-0064
NOTE 30p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *Criterion Referenced Tests; Item Analysis; *Measurement Objectives; Student Evaluation; Student Placement; *Test Construction; *Test Use; Test Validity
IDENTIFIERS *Test Specifications

ABSTRACT

Three kinds of instructionally sensitive achievement tests are described which provide useful information on formal schooling proficiencies: placement, progress, and attainment tests. Procedures to design, develop, and empirically verify attainment and placement tests are presented. The methodology is designed to ensure that the test instruments and results serve carefully defined functions and accurately describe and reflect instructional program effects. The specific concepts and skills addressed and the emphasis they receive in the instructional materials and procedures provide the basis for defining the test and reporting structure. The attainment test development process involves: (1) identifying and analyzing instructional segments; (2) specifying, and representing accurately and proportionately, the skills, concepts and outcome for each segment; and (3) verifying generated prototypical items by analyzing proficiency patterns of prototype test results. After attainment test preparation, placement tests can be developed. Items selected should yield information to differentiate student assignment to the most appropriate initial instructional segment. Preliminary results indicate that the methodology is extendable to a broad range of instructional programs and product systems. (Author/BS)

SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT

Development and Verification of Instructionally Sensitive Achievement Tests

TR 60 December 15, 1986

SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT

TECHNICAL REPORT 69

December 15, 1980

DEVELOPMENT AND VERIFICATION OF INSTRUCTIONALLY SENSITIVE ACHIEVEMENT TESTS

Ralph A. Hanson, George E. Behr, Barbara T. Meguro, and Jerry D. Bailey

ABSTRACT

Three kinds of instructionally sensitive achievement tests are described which provide useful information on the proficiencies addressed by formal schooling: placement, progress, and attainment tests. Procedures to design, develop, and empirically verify such tests are presented.

DEVELOPMENT AND VERIFICATION OF INSTRUCTIONALLY SENSITIVE ACHIEVEMENT TESTS

Ralph A. Hanson, George E. Behr, Barbara T. Meguro, and Jerry D. Bailey

From the early 1930's (e.g., Tyler, 1934) to contemporary times it has been regularly acknowledged that the standard technology for developing achievement tests yields measures that are insensitive tools for measuring instructional program effects (Tyler, 1972; Buros, 1977; Hanson, Schutz & Bailey, 1980; Madaus, Airasian & Kelleghan, 1980). Insensitive in this context means they are inadequate for identifying instructional effects and exemplary schooling practices (Hanson & Schutz, 1978). However, such instruments continue to be developed and used at least in part because there are seemingly "no alternatives" (Buros, 1978).

A methodology for providing instructionally sensitive tests that has been formulated, tested, and replicated in practice is presented in this report. It entails three kinds of achievement tests, each of which has clearly defined information functions in connection with an instructional product system.

Context

The report focuses on the method for developing instruments rather than on the broader methodological context within which the instrumentation technology was derived and verified. Background on this broader methodological context may be found elsewhere (Hanson & Schutz, 1978; Hanson, Bailey & Molina, 1980). However, it is relevant to note that this context is termed programmatic educational R&D and has been nurtured over the past decade and a half in various forms by Regional Educational Laboratories and R&D Centers.

One of the important outcomes yielded by work in the Laboratories and Centers during the late 1960's and early 1970's was the development and implementation of instructional product systems. These new product systems appear at first look to be simply "more instructional materials." However, they differ in a number of ways from conventional instructional materials. For the most part, these differences are in degree rather than kind, which make them unobtrusive. For example, the design specifications, which are the blueprints for research-based instructional product systems, are derived from careful analytical and empirical inquiry rather than tradition, the "consensus" of curriculum experts . . . etc. Similar differences can be found in the way actual instructional materials

are prepared and tested and the way personnel training and installation components are developed. The latter components provide direct support for school efforts to use the product system.

Programmatic R&D efforts at SWRL have contributed several comprehensive product systems for instructional use in schools, permitting a reliability of schooling effects not previously available (e.g., ... Hanson & Schutz, 1978, Hanson, Bailey & Molina, 1980; Hanson, Schutz & Bailey, 1980). Reliability of effects simply means that when these product systems are used in schools under usual conditions, defined instructional outcomes are attained with less variance and higher replicability than with other forms of instruction. Furthermore, the variance observed in effects can be linked directly to the operational practices employed in the use of the product systems.

While such product systems have obvious value to schools and to educational practice in general, they also provide the basis for a new kind of research effort. This research effort centers around the use of product systems as the instrumentation system for studying major educational issues. One such issue is achievement testing in schools, and the methodology described here was derived from single- and multi-year inquiries pertinent to this issue using various product systems and conducted with the cooperation of many school districts across the country.

INSTRUCTIONAL ACHIEVEMENT TESTS

In this section of the paper, the characteristics and specific functions of the instruments which are yielded by the method are described. Subsequent sections describe procedures for constructing and verifying these instruments.

Three specific kinds of tests are treated here as necessary and sufficient for describing achievement in connection with an instructional product system. These are referred to as placement, progress and attainment tests, and together they constitute the measurement elements of an instructionally sensitive instrumentation system. Descriptive characteristics of each kind of test are given in Table 1.

Placement Tests

Placement tests provide information that is used to guide the instructional assignment of students prior to involvement of a student in a given instructional program. This information can be used to help select students who can benefit from the instruction and to identify a segment of the product system where the student might best begin work. Another use of placement test information is

Table 1: Characteristics of Tests Forming an Instructionally Sensitive Instrumentation System

Test	Instructional Unit Referenced	Instructional Time (Hrs.) Referenced	Schooling Boundaries	Score Referents	Expected Results	Typical Reporting	When Given
Placement	A set of related instructional segments	60 - 120 (2-4 segments)	1-6 years	single segment	consistent with structural relationship between segments	needs instruction/does not need instruction	before instruction begins
Progress	A topic or unit of a segment	5 - 10	1-6 weeks	unit or topics in a unit within a segment	high	proficient/ non-proficient	during instruction
Attainment	A single segment	30 - 50	1-6 months	full segment or outcome areas of a segment	high	continuous (percent scores)	after completion of a segment

to provide a description of the skills/concepts of a student or of student groups in a given instructional program. Such descriptions can be useful as baseline information for evaluating program effects (Hanson, 1980).

Selection, placement, and baseline information have conventionally been derived from standardized achievement test scores and teacher judgments. However, evidence gathered through product system exercises shows that such "laisse-faire" approaches to pupil selection/placement can result in significant losses in school effectiveness, especially due to underplacement of students (e.g., Hanson, Bailey, & Molina, 1980; Behr & Hanson, 1977).

Progress Tests

Progress tests serve to provide information on a student's learning status during the course of instruction. Such tests are used at frequent intervals (often daily or weekly, and typically within monthly intervals). The information provided serves as the basis for the immediate assignment of instruction. Also, it provides a timely indicator of student progress in terms of lessons completed. Aggregates of this information for classes and schools yield fine-grain information on rate and amount of instruction completed and as such serve as markers of product system implementation. Other than in aggregated form, progress test information has little value for audiences outside the classroom since it addresses instructional management rather than pupil attainment.

Inquiries carried out using product systems have verified these points and provided some insights into issues surrounding progress tests. One specific finding is that progress tests need not be referenced to a single "objective," a procedure which until recently had been widely advocated (Popham, 1975; Wolf, 1979). Put another way, the frequency and precision of progress test information suitable for self-instructional programs is far greater than the information function such tests can reasonably perform for instruction in conventional classroom settings (Follettie, 1980).

Another related finding is that the methods used to obtain progress test information can often be integrated into instructional activities making them virtually unobtrusive. While formal progress tests may serve well in the context of the typical self-instructional sequence, they are not universally appropriate for classroom-centered instruction. Where students are performing instructional tasks on a regular basis, formal extrinsic progress tests are unnecessary and undesirable.

Attainment Tests

Attainment tests serve several functions. One function is to acknowledge that notable student learning has (or has not) occurred in the instructional program. This achievement is reflected in the proficiency displayed on attainment tests. Alternately stated, the acknowledgment function clearly and concretely describes what students do and can be expected to learn in instruction that entails the product; an attainment test is the operationalization of the direct effects of an instructional program.

A related function is to serve as an "output" measure for program "evaluation" and communications. For the communication purpose, attainment test proficiencies are usually aggregated by classes or schools. They are then used both to describe overall (aggregate) effectiveness and as a dependent variable for research aimed at identifying the factors contributing to attainment (e.g. Hanson & Schutz 1980; Hanson, Bailey & Molina 1980) The research information when properly assembled can serve as an operational basis for instructional planning (Hanson, 1978).

These functions of attainment tests are not fulfilled by tests typically used in school settings. Standardized achievement tests do provide indicators of general learning with little relationship to either instruction received or product system effects (Hanson, Schutz & Bailey, 1980; Madaus et al., 1980). Teachers or district R&D staff sometimes provide a form of attainment test referenced to "instructional objectives." These instruments usually do not provide adequate information about instructional effects from either a descriptive or planning perspective. Publishers and other suppliers of instructional products also provide tests. However, the instruments often turn out to be progress tests rather than attainment tests and thus are not able to fulfill the descriptive and planning information functions of attainment tests.

Chronological Test Development Schedule

The three kinds of tests reference related aspects of an instructional program and therefore are interdependent in design and use. In real-time operational use with a product system, placement tests come first, progress tests second, and attainment tests last. However, this is not the optimum chronology for design/development activities. In generating the tests, the progress tests emerge first as the development of instructional segments of the product system is completed. The progress tests operationalize the outcomes (i.e., skills/information) being taught in the specific activities to which it refers. They should not include either outcomes taught earlier or outcomes taught in a different form than presented in the instruction referenced.

The second test development effort focuses on attainment tests and can reasonably begin only after development for at least one segment of a product system has been completed. Operationally this usually means that all progress tests (or prototypes of them) would by this time be available for the segment. With this chronology, the instructional specifications prepared for the attainment tests can serve as important analysis/verification for the instructional design/development effort. To fully complete the construction of attainment tests for a product system, it is necessary to have available all instructional segments and accompanying progress tests. The separate instructional specifications and accompanying test specifications can then be checked for consistency and overlap before proceeding further.

The development of the placement test must await the development of all attainment tests since it requires the use of both the attainment test specifications and the empirical verification data on them. The placement test is prepared by selecting items from the completed attainment tests using both data and specifications. Individual items which best differentiate pupils completing one segment from those completing the next segment are selected.

Reasonable procedures for designing and developing progress tests are available in the literature of test construction and self-instructional technology and so require no additional elaboration here. In the following sections, specific procedures for developing and empirically verifying attainment and placement tests are presented and discussed. It is assumed that both the instructional materials and progress tests for the product system are completed and available.

ATTAINMENT TEST DEVELOPMENT

The process of preparing attainment tests typically takes place in three phases; instructional specifications, test specifications, and test verification. A brief description of the major tasks in each phase is given in Table 2.

Instructional Specifications

There are two major tasks in this phase. The first is to structure the segments of instruction to be dealt with. As indicated in Table 1, it is recommended that each attainment test is designed to measure a segment of 30 to 50 hours of instruction. This segmentation pattern is based on several considerations. Given current educational practices, it corresponds roughly to a quarter or a semester of instruction in a subject area for a class. More importantly, it approximates the minimal amount of instructional

TABLE 2

DESCRIPTION OF PHASES IN PRODUCING INSTRUCTIONAL ATTAINMENT TESTS

Phase	Major Tasks	Product
Instructional Analysis	1. Specify the instructional segment to be assessed 2. List the skills/concepts taught. For each: a. List or define the elements practiced b. List the practice format c. Determine the amount of practice	Description of the skills/concepts taught by instructional segment
Test Construction	1. Determine the skills and concepts assessed 2. Designate the item format for each skill/concept 3. Specify boundaries for each skill/concept 4. Specify the item sampling plan based instructional emphasis	Preliminary test specifications and prototype tests
Test Verification	1. Generate prototype item sets using test specifications 2. Distribute to users 3. Score tests and analysis of results 4. Identify actual test items	Final test specifications and final tests

time for educational effects to occur that have meaning for audiences outside the classroom (see e.g., Tyler, 1934), which is the prime audience for attainment tests.

The second major task is to analyze each instructional segment to identify the skills/concepts presented and the amount of direct instruction provided on each. During this analysis, several aspects of each instructional element (i.e., skill/concept) should be noted. These are conveniently described and illustrated via an example of such specifications. Sample instructional segment specifications are given in Table 3 for one segment (Block 8) of the SWRL/Ginn Reading Program. The attainment test was structured to provide separate scores on two outcomes areas entitled Word and Sentence Meaning and Paragraph and Text Interpretation.

1. Format designation. For each element (skill/concept) taught in a segment, the specific characteristics of the way it is practiced during instruction are noted. Thus, For the Word and Sentence Meaning outcome areas described in Table 3, students learn specific words using sentences with a multiple choice format, with an average syntax value of 256, that have no new words in the stem, and with new words used in the foils.

   For the second outcome in Table 3, Paragraph and Text Interpretation, these same specifications apply plus others associated with the various types of question. As the note indicates, examples of each question type are included in the actual specification. Here just the type of question is listed.

2. Element designation. The elements referenced by a specific format are to be listed and described. Thus, for the Word and Sentence outcome area in Table 3, this set is defined by a list of words that could be the object of a question. For Paragraph and Text Interpretation the elements are paragraphs used in instruction, defined in terms of length, range of acceptable syntax complexity, and specific vocabulary.

3. Identification of subsets of elements. The amount of practice provided for each cluster of elements is determined by counts of the frequency of practice. Using this information, categories corresponding to various different levels of element emphasis within a segment can be ascertained. Eventually, interest will center on those elements emphasized sufficiently to be considered taught to most students. These elements (or a sample of them) will eventually be included on the prototype test forms.

TABLE 3: SAMPLE INSTRUCTIONAL SPECIFICATIONS[1]

SWRL/Ginn Reading Program - Block 8

Skill or Concept	Format	Stimulus Characteristics	Elements Practiced	Frequency
Word and Sentence Meaning	Sentence Completion; Multiple Choice	Average Length of Sentence: 7 words Average Syntax Value:[2] 2.56 Average Number of Other New Program Words in Stem: 0	Storybook Words: Practiced both in stories and in workbook activities	268 words are taught[4]
			Non-Storybook Words: Practiced only in workbook activities	209 words are taught[4]
Paragraph and Text Interpretation	A Passage Followed by Multiple-Choice Questions	Average Syntax Value:[2] 2.58	Literal Questions[3]	66 items
			Concept identification in the question[3]	5 items
			Concept identification in the answer[3]	14 items
			Title/Main Idea[3]	8 items
			Purpose[3]	10 items

[1]Information in this table is taken from Final Block Assessments for Elementary CSP, a deliverable under Task 1.5.2 of N.I.E. Contract No. NE-C-00-3-0064, SWRL Educational Research and Development, Los Alamitos, CA, May, 1977.

[2]Botel, M. and Granowsky, A. A formula for measuring syntactic complexity: A directional effort. Elementary English, 1972, 49 (April), 513-516.

[3]Definitions and examples are included in the complete specifications but are not reprinted here.

[4]Exact word lists are included in the complete specifications but are not reprinted here.

Test Specifications

Once the instructional analysis has been completed for each segment of the product system, the process shifts to the second phase, test specifications. The intent here is to specify the characteristics of those subsets of the instructional element clusters that would be expected to be learned by all pupils completing the segment, i.e., to eliminate elements that did not receive enough attention in instruction to be learned. Each cluster so identified must then be represented accurately and proportionately as part of the test specifications. The following activties need to be carried out:

1. Identify anticipatory skills and concepts in segments. These are the elements that are subsumed in segment in anticipation of learning in subsequent segments and should not be included in test specifications. Since the purpose of the test is to describe the instructional attainment of students, classes, and schools there is no reason to assess anything but direct effects of instruction, i.e., those skills and concept that would be learned upon the completion of an instructional segment and represented in their most highly developed form.

2. Identify patterns of instruction emphasis across segments. Skills and concepts that are addressed in more than one instructional segment need to be identified during formulation of the test specifications. This is why the instructional specifications (phase 1) for all segments are needed before phase 2 can be completed. Depending on the instructional format and organization within segments, a given element may be taught definitively (i.e., to mastery some would say) in one segment; may be taught in part in several segments; or may never be taught to proficiency.

   Some examples of possible patterns of instructional emphasis of the same skill structure over segments are described in Table 4. Note that patterns do occur where instruction is provided on skills in segments after proficiency is expected. This instruction, however, should not be tested beyond the segment in which skill proficiency is expected.

3. Segment subscores/outcomes. Within a given instructional segment, it is unusual for more than a single score to be required to adequately measure instructional attainment (Hanson, 1980). However, it is often desirable to have two or three outcomes to adequately describe the instructional outcome attainments. The notion explicit in this statement is that a primary purpose of an "outcome area" is to provide a description of a segment of instruction at a

TABLE 4

FIVE ILLUSTRATIVE INSTRUCTIONAL PATTERNS OCCURRING ACROSS FOUR PROGRAM SEGMENTS FOR AN OUTCOME

Pattern	Description
1. N N I ▲ N	Instruction only in segment 3 with proficiency expected after segment 3.
2. I N I ▲ N	Instruction in segments 1 and 3 with proficiency expected after segment 3.
3. I N I N	Instruction in segments 1 and 3. Proficiency not expected.
4. I ▲ N I I	Instruction in segments 1, 3 and 4. Proficiency expected after segment 1.
5. I I ▲ I I	Instruction in segments 1, 2, 3 and 4. Proficiency expected after segment 2.

Legend
I - Instruction given in segment
N - Instruction not given in segment
▲ - Pointer marking when proficiency is expected and testing would take place

level of detail that makes it understandable to persons not intimately acquainted with the instructional program (i.e., those not delivering day-to-day instruction, such as administrators, parents, school board members). The highest level of generality that allows for meaningful effects to manifest themselves is soug' . This usually results in three or fewer scores per segment.

The fact that an outcome area designation may be popular jargon, e.g., reading comprehension, math problem solving, does not mean that the resulting attainment tests can be readily compared to other tests with subscores referencing the same categories, e.g., standardized tests. The defined structure is applicable in a particularized form to an instructional program. Another way of illustrating this point is by considering the tests produced via this method for two instructional programs. While they might conceivably share common outcome via descriptors, it would be unlikely that the tests would produce comparable results in use with groups of students receiving instruction in either program and, in fact, have been shown not to (Hanson & Bailey, 1980). This is because the essence of their respective scope and nature is contained in the distinctions between their respective test specifications (e.g., the form of the questions, lexicon used, and allowable syntactical structures). Such differences are usually only detectable when test specifications are carefully prepared and empirically tested (Hanson & Bailey, 1980). Often they cannot be detected even when comparing two different test forms to one another. The point is, that segment structures do not typically have and should not be interpreted as having common interpretability simply on the basis of their titles.

4. Resolve the "number of items" question. One of the most important features of attainment tests are the economy in testing time they afford over other types of achievement tests. Consistent with good general measurement procedures, multiple independent observations (to be referred to as items) are required for each outcome area (or single segment score) of an attainment test. However, the number of such observations or items required to provide the level of accuracy for the uses of attainment tests are considerably less than might be expected. Experience with such tests suggests that 30 items is the absolute maximum number required. This guideline assumes that student level score interpretation will center around distinctions in minimal proficiency (typically less than 60%), preliminary proficiency (typically 60% to 80%), and consolidated proficiency (typically 80% or more). Also, in

determining the exact number of items, the type of format employed (true-false, multiple choice, sentence completion, essay), the number of elements referenced, and more generally, the extent to which individual items discriminate between students and student groups receiving different amounts of instruction are important. The latter aspect is important because it refers to the information yield of an item. Where information yield is high, relatively few items are typically required. For example, an attainment test score may be based on as few as three or four items and function effectively.

5. Sampling of institutional skills/concepts to be tested. A strategy for sampling based on the relative instructional emphasis given to element clusters of a segment or outcome area of a segment must be devised. Frequency counts of amounts of practice provided derived from the instructional specifications can be treated like "weights" showing the relative importance of the various clusters. The sampling across strata is then determined by the amount of practice given in the instruction.

The specifications in Table 5 illustrate the results of steps 1 to 5 for the same segment (Block 8 of the SWRL/Ginn Reading Program) referred to in the earlier discussion of instructional specifications (see Table 3). The two specific outcomes, i.e., Word and Sentence Meaning and Paragraph and Text Interpretation, refer to different but complementary aspects of the instruction. While they are loosely related in that one would expect students doing well on Paragraph and Text Interpretation to do well on Word and Sentence Meaning (but not vice versa), they were differentiated in instruction by different kinds of practice. More importantly, they represent language skill areas that are often differentiated in reading tests. Thus, in spite of the fact that one skill area might be subsumed under the other, they were treated as separate outcome areas for purpose of attainment testing.

A noteworthy distinction between test specifications (Table 5) and instructional specifications (Table 3) is the sharply increased level of specificity required for the former. Instructional specifications can be (and are) more general than those for a test since not everything presented in instruction is taught and not everything taught is tested. However, the opposite condition must hold, i.e., everything tested must be taught. This is what the additional constraints of the test specifications are designed to ensure. The general guideline for these specifications is referred to as the "least common denominator" approach. It requires everything defined by the test specifications be clearly taught in the instruction, but not that everything taught be encompassed by the specifications.

TABLE 5: SAMPLE ATTAINMENT TEST BOUNDARIES[1]
SWRL/Ginn Reading Program: Block 8

Outcome	Item Format	Stimulus Characteristics	Distractor Characteristics	Content Parameters	Sampling
Word and Sentence Meaning (30 items)	Sentence Completion; Multiple Choice	1. Item stem should include only words taught prior to this block. 2. Sentence length should be about 7 words. 3. Syntax value should be about 2 or 3. 4. Sentence should not discriminate against subgroups such as black dialect.	1. All distractors should be new words taught in this block. 2. Distractors should be clearly wrong, not based on shades of meaning. 3. Distractors should be the same part of speech as the answer.	Storybook Words (weighted twice as much as non-storybook words)	Unit / Items: 1 / 5; 2 / 5; 3 / 4; 4 / 5; Total=19
				Non-Storybook Words (weighted proportionally to frequency)	Unit / Items: 1 / 1; 2 / 2; 3 / 2; 4 / 2; Total= 7
				New Words that use the same decoding skills	4
Paragraph and Text Interpretation (14 items)	Passage Followed by Multiple Choice Questions	1. Question should not be answerable without reading the passage. 2. Passage should have an average syntax value of about 2.5. 3. Passage must meet the specific characteristics for the type of item (literal, concept identification, etc.)[2]	1. Usually one distractor of each of these types: a) partially incorrect b) opposite c) plausible in reality but unrelated to text 2. All distractors must be plausible. 3. Distractors for one question should not provide clues to another question 4. Do not use "story doesn't say." 5. Do not require fine level discriminations.	Literal Questions	9
				Concept Identification in the Question	1
				Concept Identification in the Answer	1
				Title/Main Idea	1
				Purpose	1

[1] Information in this table is taken from Final Block Assessments for Elementary CSP, a deliverable under Task 1.5.2. of N.I.E. Contract No. NE-C-00-3-0064, SWRL Educational Research and Development, Los Alamitos, CA, May 1977.

[2] Specific requirements of these types of items are included in the full domain boundaries but are not reprinted here.

An example to illustrate this point can be seen in the first stimulus characteristic for Word and Sentence Meaning in Table 5. It states that item questions should include only words taught in earlier segments (i.e., Blocks) of instruction. In the actual instruction on this block, some words in item questions from the current segment were used. However, use of the current block words in items would directly confound attainment measurement of the Word and Sentence outcome area since each item would not measure the meaning of new words in isolation. Thus, the test specifications are more restrictive than those actually used in instruction.

A variation of "least common denominator" approach is applied to the second stimulus characteristic in Table 5. It states that the length of the stimulus (stem question) should be about seven words. This length is the median value found in the instructional materials.

The specifications in Table 5 also indicate the pedagogical categories and sampling to be carried out for the test. The pedagogical categories for the Word and Sentence Meaning outcome include three different categories of words taught in the segment. These are enumerated during the instructional analysis when the relative amount of practice given to the three kinds of words is specified. Only those words practiced enough to be taught are included. These words are then sampled by strata to produce the final set of concepts to be tested for the segment.

Test Verification

The boundaries provided in the Test Specifications phase are fully sufficient as a basis for generating prototypical items for each strata of a segment outcome. A set of these items, typically larger than the number of items actually used for the test, is prepared and distributed to instructional program participants for tryout. Often several forms of a prototype test are prepared to obtain data on items.

The purpose of the tryout is to obtain item data from student/classes completing various portions of each instructional segment. The data so gathered are used to revise test specifications and to select the specific items to be included in the completed attainment test. These data are also used in selecting items for the placement test which will be discussed in the next section.

The item data are used in several ways in attainment test construction. The first use is to ascertain the instructional sensitivity of items. Because of the nature of the segment definitions, there should be a clear pattern of increasing item proficiency by pupils completing more segments of instruction. This

statement applies both to items and to composites of them forming the outcome area and segment scores. Further, these results should hold across all units of analysis, i.e., students, classes, schools, and districts. Any exceptions to this pattern are reasons for careful examination of both the instructional specifications and the outcome area boundaries.

Some examples of the kind of results these kind of data provide are given in Table 6. The table shows results for items displaying six different patterns (labeled a to f) of proficiency change across instructional completion quartiles. The quartiles correspond to the division of actual class level data into four groups based on the amount of instruction completed during a school year. Note that patterns a and b show the desired profile of regularly increasing proficiency with increases in instructional completion. On the other hand, patterns c and d show profiles that do not follow the expected pattern. Pattern c shows essentially "no change" across quartiles and pattern d of alternately increasing-decreasing proficiency.

The data presented in Table 6 for patterns e and f respectively provide examples of undesirably high pre-instructional proficiency and undesirably low post-instructional proficiency. Patterns a and b both show desired pre- and post-instructional proficiency levels for items.

Some of the major reasons an item may not perform as expected are listed below and may be used to guide revisions.

1. Technical flaws. These might be due to unclear directions, misleading foils, and misinterpretation of question.

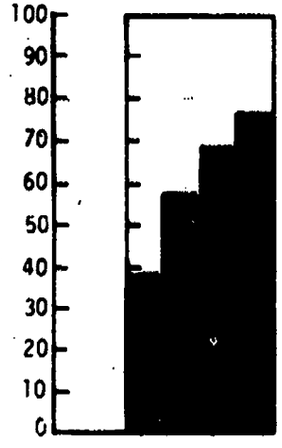
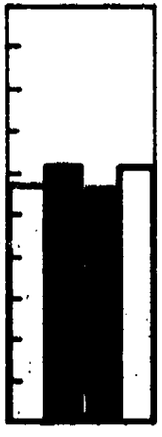
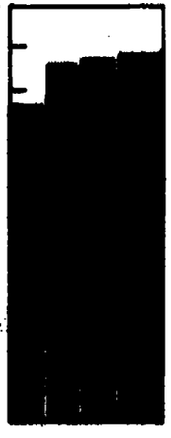
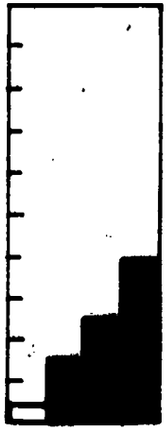
2. Inappropriate assignment to segment. This involves faulty indexing to an instructional segment so that students either learn it earlier (high proficiency pre and post) or later in the program sequence (proficiency is too low for students completing the instruction).

3. Inappropriate pedagogical referents. The item requires skills/concepts not provided in the instruction. When this happens, the specifications usually need to be revised.

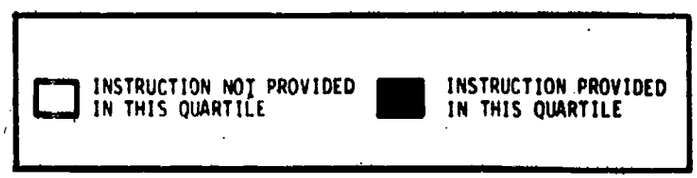
Preparation of an attainment test uses the verification data as the basis for identifying items in a quantity indicated by the test specifications. The attainment test for each segment will thus be composed of items that are sensitive to instruction in proportion to their emphasis in instruction.

TABLE 6

Proficiency Patterns Related to Instruction Received[1]

Desirable Proficiency Patterns		Undesirable Proficiency Patterns			
a. When outcome is taught during all four quartiles	b. When outcome is taught during some quartiles	c. No change in proficiency	d. Alternately increasing and decreasing proficiency	e. Initial proficiency too high	f. Final proficiency too low
WORDS	(HYPOTHETICAL)	SEQUENCE	WORD RECOGNITION	LETTER NAMES	WORD ATTACK
					
Completion Quartile 1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
Mean % 39 57 68 75	18 54 56 84	57 61 56 61	77 39 55 53	74 85 86 88	3 15 25 40
Number of Classes 21 57 54 84	— — — —	15 17 49 159	2 4 3 2	21 57 54 84	21 58 53 66

[1]All figures on this page except "b" display proficiencies attained by pupils on the outcomes of various kindergarten reading programs, as reported by Hanson, R. A., Schutz, R. E., and Bailey, J. D., in Program-Fair Evaluation of Instructional Programs: Initial Results of the Kindergarten Reading Readiness Inquiry, Technical Report 57, SWRL Educational Research and Development, Los Alamitos, California, 1977, pages 33, 38, 40, and 43. Figure "b" gives hypothetical data since none of the outcomes displayed this pattern.



PLACEMENT TEST DEVELOPMENT

Given that attainment tests have been prepared and verified for each segment of a product system, placement test development can take place. The essential task in preparing a placement test is to select items that yield information to differentiate student assignment to the most appropriate initial segment of instruction. To select these items, data must be available and used in conjunction with the attainment test specifications.

Placement Test Item Selection

What is ideally sought is a small set of items per segment that show direct change from pre to post on instruction, yet are relatively independent of instruction received in adjoining segments. The kind of information used for this purpose is simply the average proficiency of a sample of students on attainment test items from several segments after completing one or more instructional segments. Such data show how performance on an item changes with the completion of various instructional segments.

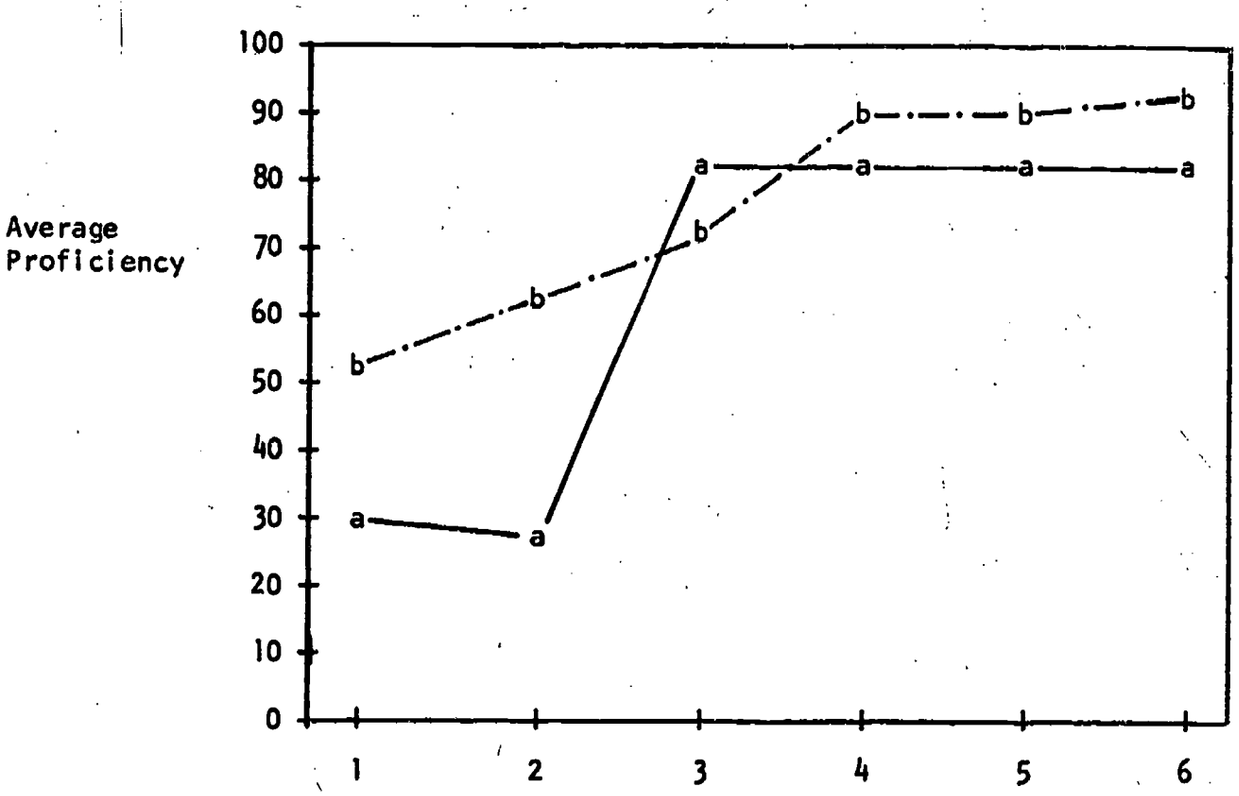
The results presented in Table 7 illustrate how such item data actually appear and are used. It presents the average proficiency of samples of students who have received instruction in various segments on two items (a and b).

Item a shows the pattern of proficiency change that is sought in placement test items. Students not completing instruction in segment 3 attain low levels of proficiency on this item. Those receiving instruction on this segment (and subsequent segments) attain high levels of proficiency. Items like a will typically measure skills/concepts that are relatively specific to the segment they reference (in this case segment 3).

Skills/concepts, that are taught across several segments typically show some sensitivity across several segments and hence are not efficient for a placement test. Item b in table 7 illustrates how data on such an item typically appears. Proficiency increases gradually for student groups completing segments 1, 2, 3, and 4 and remains at high levels thereafter. Such items are not the most efficient for use on a placement test.

To summarize, the items picked for a placement test should be those which the instructional specifications indicate are taught exclusively in a segment and the proficiency data indicate clearly differentiate student groups completing from those not completing the segment.

Table 7. Illustrative Data for Selecting Placement Test Items



Item		Segments					
a	——	28*	26	80	80	80	80
b	—·—	50	60	70	88	88	90

*NOTE: Each value is an average proficiency based on 25 or more students who have completed the instructional segment.

Placement Test Assembly

The composite placement test is made up of item sets corresponding to each segment. Typically the number of items required for each segment is small, i.e., 8 to 10. Thus the full placement test for an instructional product system with six segments is typically less than 60 items.

To interpret the results of placement test use, a "cutoff" for each set of segment items is needed. The cutoff is simply a single number guide for rule-of-thumb use by those responsible for a student's initial instructional program placement. To obtain the cutoff scores, data on the proficiency level attained on the placement items by student groups who completed each segment are used. The cutoffs are derived by simply adding up the average difficulties on the placement test items for each segment and rounding to the nearest whole number. In practice, this usually means a student must attain 7 or 8 right out of 10 in order to be credited for completing a segment for placement information purposes.

Placement Test Verification

Verification of the operational effectiveness of a placement test can be examined using either item or score level data. However, assuming the item level data used for selecting items were based on reasonable-size and representative samples (e.g., at least several hundred students from a variety of schools), the primary focus in empirical verification should be on score level data. One kind of data which is relatively easy to obtain are the placement test subscores of students corresponding to each segment and scored 1 or 0, i.e., pass/fail, based on a designated cutoff score.

Verification using such data focuses on answering the following question: Are the placement patterns observed consistent with the structure of the instructional materials? The typical expectation is that student placement patterns will display a form of Guttman scale, i.e., patterns should <u>not</u> show reversals. For example, if a student exceeds the cutoff score on segment 3, the student should also exceed this level on segments 1 and 2.

An example of such verification data is given in Table 8. It is based on data from over 8,000 students from several districts on a placement test referenced to a reading program with eight segments. The data show that the expected placement patterns were observed overall for about 90% of the 8,208 pupils receiving the test.

TABLE 8

Summary of Placement Patterns

Segment	Expected Patterns 1	2	3	4	5	6	7	8	Number placed in segment	Number of Students with Expected Pattern	Percent of Reversals
1	-								1,350	1,248	8
2	+	-							1,062	937	12
3	+	+	-						658	574	13
4	+	+	+	-					629	598	5
5	+	+	+	+	-				1,052	885	16
6	+	+	+	+	+	-			613	522	15
7	+	+	+	+	+	+	-		697	611	12
8	+	+	+	+	+	+	+	-	761	761	---
TOTALS									8,208	7,522	10

+ indicates above cutoff for the segment

- indicates below cutoff for the segment

SUMMARY

This paper discussed a framework for achievement testing in instructional programs that have identifiable intentions and resources. The framework entails three kinds of tests: placement, progress, and an attainment. A precise method for designing and developing the instruments was then presented. The methodology is designed to ensure that the test instruments and results serve carefully defined functions and accurately describe and reflect instructional program effects. As such, the specific concepts and skills addressed and the emphasis they receive in the instructional materials and procedures provide the basis for defining the test and reporting structure.

The central element in this framework is the attainment test and the key design feature of this instrument is the program segment. A program segment is somewhat akin to a well defined "domain" in criterion-referenced testing (e.g., Millman, 1974). However, unlike domain-referenced tests, the segment attainment test will likely include range of concepts and skills that would be regarded as heterogeneous from a domain-referenced test perspective. The logic for including such items within the same test (and perhaps the same score) resides in the architecture of the instruction and reporting information rather than in domain logic. The major issue in determining whether multiple scores are appropriate in an attainment test is the diversity present in terms of the instructional formats used and the relative ease with which the attainments can be described to audiences outside the classroom. These concerns often converge in practice, i.e., instructional structures that use different formats usually require multiple scores to describe the effects.

The method described in the paper is clearly appropriate in connection with any instructional product system used in a formal schooling instructional program. Preliminary results indicate the methodology is extendable to a broad range of instructional programs and product systems (Hanson & Bailey, 1980).

References

Behr, G., & Hanson, R. A. Differential access to instruction; a source of educational inequality. Paper presented at AERA Annual Meeting, New York, New York, April, 1977.

Buros, O. K. Fifty years in testing: Some reminiscences, criticism, and suggestions. Educational Researcher, 1977, 6(7), 9-15.

Buros, O. K. (Ed.) The Eighth Mental Measurements Yearbook. Highland Park, NJ: Gryphon Press, 1978.

Follettie, J. F. Task analysis and synthesis as precursors of productive instruction. SWRL Educational Research and Development, Los Alamitos, California.

Hanson, R. A. Bringing about basic changes in education. Paper presented at AERA Annual Meeting, Toronto, Canada, April, 1978.

Hanson, R. A., & Schutz, R. E. A new look at schooling effects from programmatic research and development, Making Change Happen, edited by D. Mann. New York, New York: Teachers College Press, 1978, 120-149.

Hanson, R. A., Bailey, J. D., & Moline, H. M. The implications of intra-program placement decisions for the understanding and improvement of schooling. SWRL Educational Research and Development, Los Alamitos, California, 1980.

Hanson, R. A. Evaluation and planning. SWRL Educational Research and Development, Los Alamitos, California, 1980.

Hanson, R. A., Bailey, J. D., & Schutz, R. E. Program-fair evaluation of instructional programs: Initial results of the kindergarten reading readiness inquiry, 1977. Technical Report No. 57. SWRL Educational Research and Development, Los Alamitos, California.

Hanson, R. A., Schutz, R. E., & Bailey, J. D. What makes achievement tests tick: Alternative instrumentation for instructional program evaluation. SWRL Educational Research and Development, Los Alamitos, California, 1980.

Hanson, R. A., & Bailey, J. D. Program fair assessment, identification of program effects and empirical curriculum inquiry. SWRL Educational Research and Development, Los Alamitos, California, 1980.

Madaus, G. F. Airasian, P. W., & Kelloghan, T. School Effectiveness: A Reassessment of the Evidence. McGraw-Hill, New York, 1980.

Millman, J. Criterion-referenced measurement. Evaluation in Education: Current Applications, edited by W. James Popham, Berkeley,CA: McCutchan, 1974.

Popham, W. J. Educational Evaluation. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

Tyler, R. W. Constructing Achievement Tests, Bureau of Educational Research, Ohio State University, 1934.

Wolf, R. M. Evaluation in Education. New York: Praeger Publishers, 1979.