DOCUMENT RESUME

ED 279 682 TM 870 069

AUTHOR Herman, Joan L.
TITLE What Do the Test Scores Really Mean? Critical Issues in Test Design.
PUB DATE 86
NOTE 13p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Content Validity; *Criterion Referenced Tests; Educational Assessment; Educational Testing; Elementary Secondary Education; Measurement Objectives; Multiple Choice Tests; *National Surveys; Test Construction; *Testing Problems; Testing Programs; *Test Interpretation; Test Results; Test Validity
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

Issues in designing valid tests for the National Assessment of Educational Progress (NAEP) are discussed. Test scores are often provided without any information on the nature of the tasks represented. Because test domains are defined by individual item writers, the generalizability between tests and items is suspect. While typical content validation procedures help assure that the included items are important, they still might not represent the full range of knowledge and skills constituting given domains. As a result, the underlying meaning of what is tested is vague, and the specific definition of what is to be tested escapes public scrutiny. This is especially important when matching particular tests and curricula among states. Better specification of test content and task structure is recommended. Elements in good task structure should include: task description; content limits; linguistic features; cognitive complexity; and format. Recent NAEP assessments defined four different types of context for test items: (1) scientific, (2) personal, (3) societal, and (4) technological. Three levels of cognitive complexity items were defined: (1) knows, (2) uses, and (3) integrates. Six categories of subject content were specified. In conclusion, NAEP planners should emphasize content validity; define more specifically what is to be tested; provide better models for item construction; and assure that the entire domain is represented. (GDC)

What Do the Test Scores Really Mean?

Critical Issues in Test Design

Joan L. Herman

University of California-Los Angeles

Paper oommissioned by

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

DRAFT

WHAT DO THE TEST SCORES REALLY MEAN? CRITICAL ISSUES IN TEST DESIGN

Joan L. Herman
UCLA Center for the Study of Evaluation

The National Assessment of Educational Progress aims to provide information to the public, legislators, educators and others about students' level of performance on a broad spectrum of significant, age appropriate knowledge and skills within a particular subject area. How well are students performing? What are they able to do? Has the level of their performance changed for the better or the worse? The problem is essentially a descriptive one, with the caveat that descriptions over time probably are equally of interest. Unfortunately, however, the descriptive questions of what students are able to do and of at what level they are able to perform cannot be sensibly answered without knowing the nature of what they are being asked to do and without rigorous assurance that the items used to describe their performance adequately represent the knowledge or skill of interest. This observation, while axiomatic and longstanding, has been has not received adequate attention by those in the testing and measurement community. Researchers and test developers are quick to warn that test scores represent only estimates of a students' skills, and they have developed sophisticated models and elegant techniques to make those estimates more empirically precise and/or efficient. Their techniques, however, assume a well defined test domain, an assumption which frequently is violated raising very basic questions about what is being estimated and what a test score means with regard to the quality and level of student performance. This paper argues that NAEP's search for empirical precision needs to be matched by equal concern for conceptual precision in the specification of test content. It begins with a discussion of problems which arise when test content is not well specified, considers issues in assuring that NAEP tap the most significant subject areas skills, and recommends test specification solutions that are based in current research in cognition and the structure of expertise and that expand response alternatives beyond selected response options.

The Problem

The essential problem is this: What sense can be made out of a test score when we do not know the nature of the task it represents? Or stated alternatively, how can one validly and reliably measure some knowledge or skill without knowing the nature of the domain which the measurements are supposed to represent? The answer, according the common test development practice, is that we do not need to have a very detailed view of

what we want to assess in order to assess it well and make sensible interpretations of the results. Rather than starting apriori with a detailed, well grounded conception, we arrive at one posthoc. Consider the typical process: teachers, content experts, and/or others are assembled to generate large numbers of test items in response to a very general content process matrix; the items so generated are subjected to both empirical and judgmental procedures; the surviving items, those which are judged representative of something important and which are empirically coherent, are then assumed to adequately define the domain of interest. We leave it to the item writers and to the test items themselves, in short, to defacto define the domain.

The problem with such a process is that its base is essentially arbitrary. It aggregates the content biases and idiosyncracies of individual item writers and assumes that somehow by combining a great number and variety of individual decisions, we are left with a sound, representative domain and a set of generalizable measures of that domain. The error of this assumption is evident in a number of studies that have conducted comparative analyses of standardized test content (Herman and Cabello, 1984;Floden et al, 1980 Schmidt, 1983). These studies have found that although there is broad agreement in the various tests on the subscales that are used to constitute the assessment in each basic skill area, there is considerable disparity in the specific skills which are used to represent each subscale. Thus, for example, most standardized tests purport to measure and provide reports on students' performance in something akin to vocabulary and reading comprehension within reading; math concepts, computation, and problem solving within math; capitalization, punctuation, useage, and spelling within language arts; but the the types of items included within each scale, the relative emphases given specific skills within each area, and the specific topic coverage differs markedly from one test publisher to the next and from one state assessment to the other (Burstein et al, 1985). Consequently, the generalizability of results of one test to another, or from one set of items to another, is suspect, and the meaning of the domain inconsistent.

Gross imperfections in generalizability and the problems they pose in validly interpreting test results are highlighted when the number of items included in an assessment is small. In the extreme case, consider the NAEP writing report, Trends Across the Decade, 1974-84 (Applebee, Langer, Mullis, 1986). The report characterizes trends in student writing performance in three different genres, informative writing, persuasive writing, and imaginative writing. Because of changes in both writing prompts and in administrative procedures over the three assessment periods, responses to only one prompt per genre were available to characterize student performance at each point. Thus the meaning we can derive from such findings is directly proportionate to our confidence that each prompt adequately represents students' performance in each genre. To what extent is each prompt representative in this sense? No fationale is available, either from an empirical standpoint, e.g. evidence to suggest that

students' performance on the administered prompt fell near the midpoint of their performance on a range of tasks within the genre, or from a design standpoint, e.g., rational support, preferably research based, for the proposition that the administered prompt is modally prototypical and/or based on apriori task requirements represents some mid level of task difficulty. On the contrary, we have some evidence to suggest that students' performance within genre is not stable, and that depending on which prompt we choose to examine, we come up with significantly different pictures of student skill levels. Looking at the 1974 and 1979 assessment of nine-year olds, for example, we find that the percentage of students rated minimal or better on task accomplishment in persuasive writing in one year ranges from about 35 percent to about 75 percent depending on the prompt chosen to characterize their performance (p.45). (See figure 1) The trend data also leads to different conclusions depending on the prompts selected for scrutiny. Looking at the performance of thirteen years olds on imaginative tasks, we find that two of the three prompts show a slight upward trend from the 1974 to the 1979 assessments while the third, and the one on which the three year trend analysis is based, shifts downward over the same period (p.46). The choice of items, in short, profoundly affects performance level interpretations, and the item writer(s), not the domain itself, in many ways controls the results and their conclusions.

INSERT FIGURE 1 ABOUT HERE

While one might counterargue that averaging students' performance over a number of items, as is typically the case in multiple choice tests, alleviates some of these generalizability problems, and certainly this would strengthen the interpretability of the writing example just cited, problem(s) in content validity still remain. Consider, for example, the number of subject area topics which are supposed to be assessed by NAEP's science assessment. Any single multiple choice item typically measures only a very miniscule fraction of the specific topic, and what it specifically covers is left to the discretion of the item writer, essentially hidden from public view. Although typically employed content validation procedures help to assure that items included on the test are considered important, what assurance in there that the test items represent the full range of important relevant content? Have items been sampled broadly to be fully representative of the domain of interest or or test items concentrated in particular areas and in a constricted, but unknown, skill range? Figure 2 displays alternate pictures of how well a given number of test items covers important content within a particular topic area. Which is an accurate picture of current NAEP assessments?

INSERT Figure 2 about here

We hope, of course, for the most balanced, comprehensive picture. Returning to the example of the writing study, for instance, it would be highly desireable to sample student

performance from the full range of tasks which are typical of a particular genre in order to characterize fairly how students perform in that genre. Furthermore, in order to sample them adequately, we may well want to consider defining apriori the relevant boundaries and the varying task requirements which constitute different types of tasks within the genre. We could then be more assured that the specific exemplars selected for testing were optimally representative of the domain of interest. The problem is most obvious in production tasks where the number of items sampled is small, but also exists in multiple choice assessments featuring large numbers of items.

The tests of empiricial coherence typically employed in the development of multiple choice assessments, in fact, could mitigate against a fully balanced representation and may instead reinforce a more constricted view of a given skill or knowledge domain. The difficulty of using multiple choice items to measure deep understanding and the highest levels of cognitive skills is frequently acknowledged. Studies have also demonstrated the limits of using multiple choice items to measure higher level production skills. Early studies by UCLA's Center for the Study of Evaluation (Spooner Smith, 1980), for example, found that students' performance on multiple choice tests of writing skill did not adequately predict their actual performance in writing, even when both measures were directed at the same analytic skill categories. (Both the multiple choice test items and the scoring scheme for analyzing their writing were directed at the same elements within the domain: use of topic sentence, support, organization, usage, etc.).

Taken together i.e., the difficulty of developing test items to measure the highest levels of cognitive skill and the inadequacies of recognition items for measuring complex production tasks,, these two observations point to an important flaw in relying upon empirical coherence to validate a set of items. Within any given field test of multiple choice items, then, we might expect only a few items creatively written to assess high-order production skills and problem-solving; conversely, we might expect most of the items, because they are easier to conceive and construct, to assess lower level skills. Those items which are empirically coherent, then, may well be concentrated in lower levels of skill application and miss the most complex aspects problem solving. On the other hand, some of the items which are discarded as outliers may in fact be capturing something of real significance, critical aspects of what we're trying to measure. A constricted assessment may be the result.

In summary, there are a number of problems in the descriptive validity of test results under traditional test construction procedures: the definition of the domain rests in the hands of item writers and left to their collective biases; the generalizability of the domain definitions thus is suspect. Furthermore, while typical content valididation procedures help to assure that those things included on a test are important,

whether the test items are representative of the full range of knowledge and skill constituting given domains is moot. As a result, the underlying meaning of what is tested is slippery and the specific definition of what is to be tested escapes public scrutiny. (The very general frameworks or content/process matrices which are used to characterize test content are not the subject of the latter statement; these are quite public, but defined at a level of abstraction where few could disagree and which permit a wide variety of specific test content.)

The publicness of what is tested and the clarity and precision of its specification becomes increasingly important when the match between and among particular tests and/or curricula is an issue. For example, the equity of using NAEP in state-by-state comparisons rests at least partially in the match between the curricular intentions in each state and the NAEP items. Without knowing the underlying specific bases of the items, it is difficult to come to a meaningful determination of such a match. The problem of relying on "similar sounding" subscales for making such determinations is demonstated in the test content analyses cited above. Adding fuel to the argument is a recent study comparing subject matter contained in state assessments across the country which found great diversity in the depth and breadth of coverage on presumably similar subscales (Burstein, Baker & Aschbacher, 1985). Matches determined at this level, then, would be both superficial and artificial.

Toward a Solution

Inherent in the arguments above is a solution to the problem of more meaningful and interpretable test results: better specification of test content. This call for greater descriptive rigor is not new but harkens back to early advocacy for criterion referenced testing and later for competency tests. More recently, Baker and Herman (1983) have outlined a test design approach grounded in research in learning, instruction, and cognitive science and focused the definition of task structures. Elements specified in such structures include:

> Task description, or a general descriptor characterizing the nature of the knowledge, skill, or objective to be assessed;
>
> Content limits which circumscribe apriori the substance or content which is permissible for testing and the performance quality or level of discrimination expected, both defined by reference to the curriculum, consensus, and principles of learning and understandings of the structure of knowledge.
>
> Linguistic features, controlling the linguistic complexity of assessment so that it does not interfere with the construct actually being assessed;
>
> Cognitive complexity, or the intellectual "level" apart from content at which the items or targetted, operationalized in relation to the specified content limits;

Format, including both the descriptive modes in which the task is presented and the form in which the task is to be presented.

The task structure provides a specific, descriptive and generalizable blueprint against which test items can be generated and a public and operational statement for each domain being assessed.

Recent NAEP assessments have been moving toward such a domain specification approach and their attempts to ground definitions of skills in recent theory of cognition is commendable. But further progess is desireable, progress which could benefit not only the validity of NAEP assessments but also could provide models for local and state test development as well. Take, for example, the 1985-86 NAEP Science Assessment (NAEP, 1986). The assessment framework is a three-dimensional matrix defined by content, context, and level of cognition. The content dimension specifies six categories, including the traditional disciplines of scince, its nature and processes and its history, and specific topics to be assessed within each. The context dimension defines four different types of context for test items: scientific, personal, societal, and technological.

Perhaps most interesting is the cognition dimension which attempts to define items according to the cognitive processes required to deal with science content at different levels of cognitive complexity:

Knows: Successful performance depends on the ability to recall specific facts, concepts, principles, and methods of science; to show familiarity with scientific terminology; to recognize these basic ideas in a different context; and to translater information into other words or another format. This category generally involves a one-step cognitive process.

Uses: These exercises test the ability to combine factual knowledge with rules, formulas, and algorithms for a specified purpose. Successful performance depends on the ability to apply basic scientific facts and principles to concrete and/or unfamiliar situations; to interpret information or data using the basic ideas of the natural sciences; and to recognize relationships of concepts, facts, and principles to phenomena observed and data collected. This category generally involved a two-step cognitive process.

Integrates: These exercises test the ability to organize the component processes of problem solving and learning for the attainment of more complex goals. Successful performance depends on the ability to analyze a problem in a manner consistent with the body of scientific concepts and principles, to organize a series of logical steps, to draw

conclusions on the basis of available data, to evaluative the best procedure under specified condictions, and to employ other higher-order skills needed for reaching the solution to a problem.

This category generally inolves multi-step cognitive processes. In particular, it requires such mental processes as generalizing; hypothesizing; interpolating and extrapolating, reasoning by analogy, induction and deduction; and synthesizing and modeling. (p.10)

While this is heroic attempt to operationalize the meaning of higher, mid- and lower level cognitive skills, because the boundaries of each level are not clear it falls short of its goal of producing a useable scheme that can be used to generate and categorize test items. The differentiation between use and integrate, in particular, is often difficult to comprehend, e.g., the difference between interpreting data using the basic ideas of the natural sciences and analyzing a problem consistent with the body of scientific concepts and principles. Apparently absent also is a set of instructions, exemplars and models for generating test items for each category. Further work in clarifying and better operationalizing the meaning of each of the categories and in validating their integrity would be important contributions to both the assessment and teaching of science, providing a common vocabulary and a set of parameters that can be applied to science content and that can help focus instruction and test development. The idea is <u>not</u> to specify a common set of science objectives for all schools or all states, but rather to provide useable tools and models for stimulating the definition of state or locally sensitive goals and objectives and for generating test items covering a range of levels of cognitive complexity.

Examining the extent to which the definitions of cognitive complexity applicable to science objectives can be generalized to those in the social sciences is also worthy of exploration. Particularly at the elementary school level, a common scheme across content areas would simplify practitioners work in test development and in instruction and might contribute as well to curricular integration. At the secondary level, it might encourage communication among professionals across disciplines and contribute also to curricular integration at that level.

Coming up with more refined specifications for writing multiple choice items measuring higher levels of cognitive complexity in the various content areas would be an important steps in increasing the validity and interpretability of NAEP results, but it would not solve the problem of assuring that NAEP assesses the full range of skill development, including the highest levels of problem-solving and critical thinking. Literature cited earlier points to the limits of using multiple choice items. The ultimate goals are production skills, not recognition, which require constructed response/essay items for valid assessment.

The development of production items for assessing higher levels of problemsolving and critical thinking will require the same careful specification of task requirements and of scoring criteria advocated for multiple choice items, i.e., definition of the nature of the task that students are being asked to complete and of the attributes which control its difficulty. For example, with regard to "critical thinking," what are the defining characteristics of tasks requiring critical thinking? over what contexts --academic, personal, societal, etc. --should the tasks be drawn? At what level of complexity should the tasks be constructed? What specific content understanding or knowledge is pre-requisite to task completion? Defining the nature of a correct response and reliable, generalizable rubrics for scoring student productions is an equally important and challenging aspect of the domain specification process. What aspects or elements of the response need to be attended to? What criteria should be employed and what rules can be constructed to reliably operationalize the criteria? R&D in the assessment of writing provides models than might be translated for use in assessing content area understanding, models which are generalizable across topics areas and which yield reliable, psychometrically sound results. (See, for example, Quellmalz and Burry, 1983)

As with their multiple choice counterparts, it would be of interest to examine whether the defining charactertistics and nature of higher levels of understanding, critical thinking and problem solving tasks varies from one content area to another or whether they can be defined independent of content but applied across areas. The roots of potential solutions may well lie in research findings from cognitive psychology, artificial intelligence, the structure of knowledge, and differences in novice and skilled performance. As an example, research in both learning and in expert systems has produced protocols for representing and assessing knowledge structures (Dansereau & Holley, 1982; Novak et al., 1983; Naveh-Benjamin, 1986). To what extent can these techniques be adapted and their results quantified to provide reliable, generalizable strategies for assessing deeper levels of content understanding?

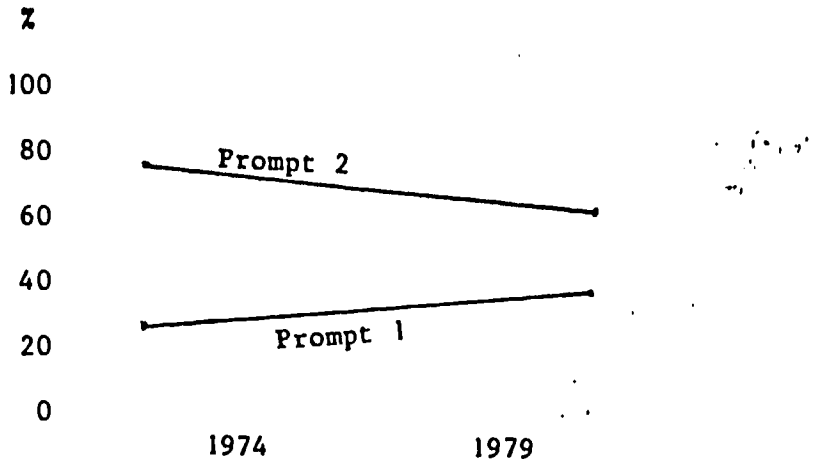
This paper has argued that NAEP planners need to give more attention to content validity issues in test design. It has recommended that NAEP define more specifically the nature of what is to be assessed, provide better model prototypes to guide item development, and institute assessment approaches to assure that the resultant assessment represent the entire domain of interest. NAEP advances in these areas would enhance the quality of the national assessment and also would provide important benefits for state and local practice.

FIGURE 1*

Differences in Performance Levels
Depending on Prompt Selection

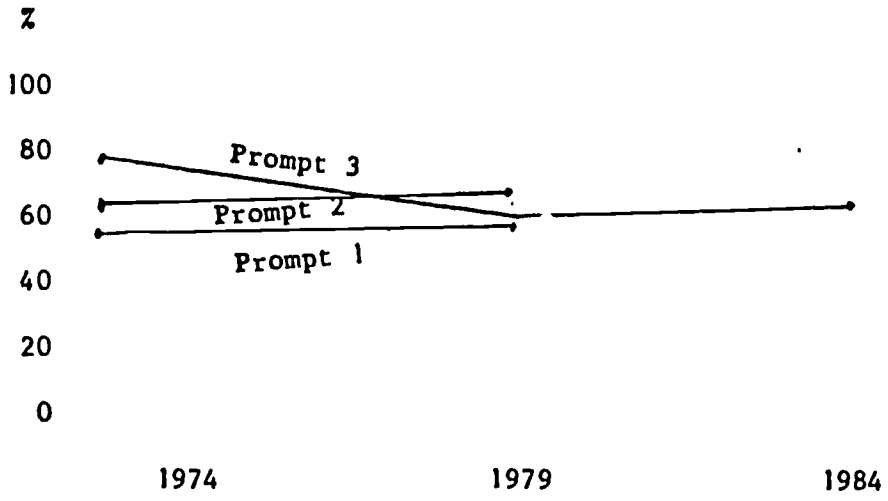
(a)

Percentage of 9-year Olds Rated Minimal
or Better on Informative Tasks



(b)

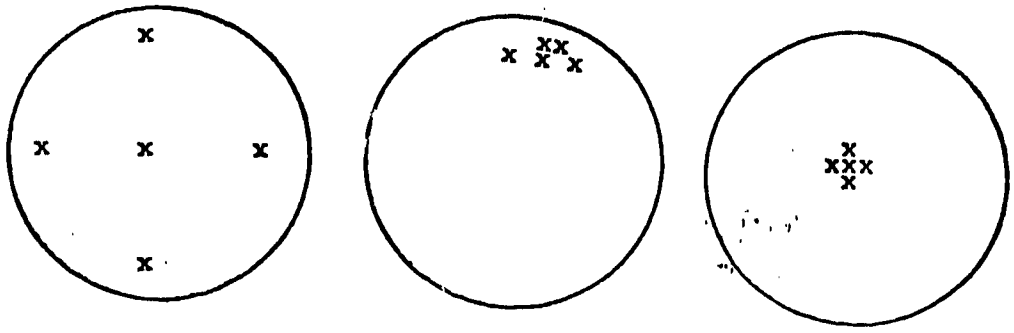
Percentage of 13-year Olds Rated Minimal
or Better on Imaginative Tasks



*Taken from NAEP, 1986, p. 45-46

FIGURE 2

Representations of Domain Coverage



References

Applebee, A.N., Langer, J.A., & Mullis, V.S. Writing. Trends across the decades, 1974-1984. Princeton, NJ: National Assessment of Educational Progress, 1986. (Report No. 15-W-01)

Baker, E.L., & Herman, J.L. Task structure design: Beyond linkage. Journal of Educational Measurement, Summer 1983, 20, 149-164.

Burstein, L., Baker, E.L., Aschbacher, P., & Keesling, J.K. Using state test data for national indicators of educational quality: A feasibility study. Los Angeles, CA: UCLA Center for the Study of Evaluation, 1985.

Dansereau, D.F., & Holley, C.D. Development and evaluation of a text mapping strategy. In A. Flammer and W. Kintsch (Eds.), Discourse Processing. Amsterdam: North-Holland Publishing Company, 1982.

Floden, R.E., Freeman, D.J., Porter, A.C., & Schmidt, W.H. Don't they all measure the same thing? Consequences of selecting standardized tests. In E. Baker and E. Quallmalz (Eds.), Design analysis and policy in testing and evaluation. Beverly Hills, CA: Sage, 1980.

Herman, J.L., & Cabello, B. Testing, evaluation and instruction: Problems in making them work together. Invited presentaion to the Management Information Network. Los Angeles, California, June 1984.

National Assessment of Educational Progress. Science objectives: 1985-86 assessment (Objectives Booklet No. 17-S-10). Princeton, NJ: NAEP, 1986.

Naveh-Benjamin, M., McKeachie, W.J., Len, J., & Tucker, D.G. Inferring students' cognitive structures and their development using the "Ordered Tree Technique." Journal of Educational Psychology, 1986, 78, 130-140.

Novak, J.D., Gowin, D.B., & Johansen, G.T. The use of concept mapping and knowledge vee mapping with junior high school students. Science Education, 1983, 67, 625-645.

Quellmalz, E., & Burry, J. Analytic scales for assessing students' expository and narrative writing skills (CSE Resource Paper No. 5). Los Angeles: UCLA Center for the Study of Evaluation, 1983.

Schmidt, W.H. Content biases in achievement tests. Journal of Educational Measurement, Summer 1983, 20, 165-178.

Spooner Smith, L. Measures of high school students' expository writing: Direct and indirect strategies (CSE Report No. 133). Los Angeles: UCLA Center for the Study of Evaluation, 1980.