DOCUMENT RESUME

ED 279 679 TM 870 066

AUTHOR Haertel, Edward H.
TITLE Domain Definition and Exercise Generation as Functions of the National Assessment of Educational Progress.
PUB DATE 27 Sep 86
NOTE 18p.; 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; *Criterion Referenced Tests; *Educational Assessment; *Educational Objectives; Educational Testing; Elementary Secondary Education; Item Banks; Latent Trait Theory; Measurement Objectives; *National Competency Tests; National Surveys; State Surveys; *Test Construction; Testing Programs; Test Items; Test Results; Test Use
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

It has been recommended that the National Assessment of Educational Progress (NAEP) specify comprehensive exercise domains to measure academic achievement, and provide a national item pool to measure the objectives in these domains. These domain specifications and item pools would serve to satisfy the increasing demand for valid, accurate, and detailed testing and interpretation. Item response theory (IRT) can be used to tailor tests to different purposes and populations. The domain would consist of a set of objectives, hierarchically organized, spanning some range of cognitive learning outcomes. The corresponding item pool would include items written according to the specifications for each objective, appropriately reviewed, field tested, and calibrated. Nationally developed domains and item pools could be used to compare states' curriculum guidelines, describe achievement trends, and quantify achievement in a manner superior to traditional norm referenced tests. NAEP, at the Federal level, provides the natural vehicle for such efforts. Domain descriptions in each curriculum area should represent the content at various levels of aggregation and reflect skill hierarchy, referent generality, and instructional sequence. Implementation issues requiring further consideration involve comprehensiveness of outcome domains, including items for all objectives and affective objectives; choice of IRT model, standards and criteria; and reporting of results. (GDC)

Domain Definition and Exercise Generation as Functions of the National Assessment of Educational Progress

Edward H. Haertel

Stanford University

Paper commissioned by

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

Domain Definition and Exercise Generation as Functions of the National Assessment of Educational Progress

Edward H. Haertel
Stanford University

Paper Commissioned by the Study Group on National Assessment, Working Group on State and Federal Roles and Responsibilities, September 1986.

Domain Definition and Exercise Generation as Functions of the National Assessment of Educational Progress

Edward H. Haertel
Stanford University

On August 5-6, 1986, the Working Group on State and Federal Roles and Responsibilities convened in San Francisco. Among the group's recommendations were that NAEP serve as a vehicle for (1) specifying comprehensive exercise domains to measure learning outcomes in school subject areas, and (2) providing a national exercise pool to measure the objectives included in these domains.

Such domain descriptions and item pools could serve as a framework for linking the national assessment with state-level assessments; coordinating State and National assessments with other subject matter content development efforts (e.g., Holmes, Carnegie, College Board); informing deliberation on achievement standards and targets in educational reform; communicating the meaning of these standards and targets; and coordinating and improving state-level curriculum planning.

This paper discusses (1) the rationale for developing common domain specifications and item pools, (2) arguments for Federal sponsorship of these activities through NAEP, (3) the form domain descriptions might take in different content areas, and (4) issues in implementation.

Need for Item Domains and Item Pools[1]

Student test scores are among the most easily understood and widely cited indicators of schooling effects, but are often subject to misinterpretation. Tests designed for one purpose are appropriated for other purposes, scores are reported in a bewildering array of noncomparable metrics, and policymakers and the public often appear unconcerned with the precise content and skills measured by the tests on which pupils, classrooms, schools, districts, and states are compared. Moreover, comparisons are often based on data from nonrepresentative samples, as when average SAT scores are reported.

The demand for valid, accurate, and detailed student performance data is further increasing, as policymakers turn more and more to tests as tools for monitoring and for influencing curriculum coverage and instructional effectiveness. The CCSSO has recently called for state-level comparisons of achievement and other school outcome measures, and increasing centralization of State educational decisionmaking is also feeding the demand for more and better data on achievement outcomes.

<u>Modeling item performance versus test performance</u>. Achievement outcomes have traditionally been defined and measured in terms of scores on intact tests. With modern psychometric methods, however, domain descriptions and item pools become preferable to intact tests for defining achievement. Unlike classical test theory, modern item response theory (IRT) takes the item, not the entire test, as the measuring unit. An item's difficulty and other statistical properties are described by a set of item parameters, conceived as fixed properties of the item regardless of the test in which it is included or the group of examinees to which it is administered. Once these item parameters are estimated, a process referred to as item calibration, different items can be used to estimate examinee scores or score distributions on the same, common scale. Some items can be released to illustrate and describe such a scale, while others are kept secure for future use. Using a single item domain, scores can be defined and tests constructed at different levels of content specificity. Likewise, tests can be focused at different levels of examinee ability. In short, IRT offers enormous flexibility in using a calibrated item pool to describe achievement scales and to construct tests tailored to different purposes and examinee populations.

<u>Domains and item pools</u>. As proposed here for the National Assessment, a domain would consist of a set of objectives, typically hierarchically organized, spanning some range of cognitive learning outcomes. Elementary and middle school mathematics might be conceived as a single domain, as might United States History. The cognitive learning outcomes of elementary science might be conceived as a single domain, whereas high school chemistry, physics, and biology might be treated as three separate domains. The objectives included in a domain might range from mastery of factual knowledge to acquisition of complex, critical skills. Corresponding to each objective would be specifications for a set of items measuring that objective, covering item format, content, and the operations to be performed by the examinee ("skill"). Illustrative items would also be provided.

The item pool corresponding to each domain would include items written according to the specifications for each objective, appropriately reviewed, field tested, and calibrated. The organization of the item pool would match that of the domain. Both the items and the domain structure could change over time, but items would be expected to change more rapidly. Knowledge is evolving quickly in some areas, and some particular items could become obsolete in a matter of a few years. Most such changes in knowledge could probably be accommodated without revising domains, objectives, or even item specifications.

Even in areas where knowledge is changing more slowly, items would have to be considered an expendable resource, and a mechanism would be required for the continuing development of new ones. The validity of items is likely to erode with repeated use, and ideally, some new items would be used in each assessment cycle to assure the assessment's integrity. Depending upon the design and purposes of the assessment, sufficient items might be required to assemble tests at different levels of difficulty, comprehensiveness, and accuracy, and to construct multiple parallel forms of these tests. Additional items spanning the range of difficulty levels for each objective might be published (as are NAEP released exercises) to communicate the meaning of each objective and of different score scale points for that objective. In some cases, it might be appropriate for teachers to use such items in their classroom instruction. Finally, items could be made available to the States or other qualified test developers, for their use in linking other tests to the NAEP scales. Such linkages might be accomplished using data from special administrations of the external test together with the calibrated NAEP items, or simply by including some calibrated NAEP items within the external test.

Uses of domain descriptions. The construction of national achievement domain descriptions and item pools would be invaluable in furthering educational reform and improvement. They would provide for the first time a comprehensive, common set of categories and metrics in which to describe student achievement outcomes. State curriculum guidelines could be compared explicitly, achievement trends could be described more accurately, and most important, achievement levels could be quantified in a manner qualitatively superior to the relative, norm-referenced scales almost universally used today.

The need for such a common set of scales has been recognized before, but has never been adequately met. Efforts to link existing intact tests, like the federally sponsored Anchor Test Study (Loret, et al., 1974), have suffered rapid obsolescence as tests were revised. NAEP, which has approached the same problem at the level of exercises rather than tests, has foundered in the massive detail of statistics on hundreds of exercises, each conceived as being of potential interest in its own right. An integrated system, permitting both fine-grained analysis of narrow learning objectives and broad, policy-relevant summaries of achievement trends, can only be developed by locating items within an organizing framework that is grounded both substantively and statistically. The 1984 NAEP reading scale (Educational Testing Service, 1985) demonstrates the potential of IRT methods to support such a system.

The set of domains envisioned would in no sense constitute a national curriculum, but it would provide a framework for describing curriculum differences. Only a subset of the objectives would be used for State-level comparisons, and there would be no need to specify these until after the domain and item pools had been developed. The domains would have to be comprehensive, representing the range of learning outcomes sought under different curricula and using different instructional approaches. They would include outcomes appropriate for students of limited ability as well as the more gifted, for students preparing to enter the world of work as well as those preparing for postsecondary education. They would also encompass a range of mastery levels, especially for complex, higher-order skills.

Comprehensive Domain Specification and the National Assessment

The creation of common domain descriptions and item pools would best be sponsored by the Federal government, and is a function well-suited to NAEP. Major functions of the domain descriptions and item pools would include both achievement monitoring at the National level and coordination of State-level assessment activities, including State comparisons. Moreover, in developing the assessment system, cooperation would be required of the States and of organizations representing interested constituencies. It would be necessary to bring together curriculum specialists, scholars in relevant academic disciplines, and psychometricians, and to assure that the interests of state and local school administrators, teachers, students, policymakers, and the public were all represented. Federal sponsorship could legitimate such an effort and could help to assure that it was carried out fairly and responsibly.

NAEP provides a natural vehicle for such a development effort. The rationale for the system is consistent with goals stated for NAEP since its inception, namely providing information on the status of, and changes in, the knowledge, skills, understandings, and attitudes of young Americans. Moreover, the National Assessment is well known and respected, and procedures have been developed within it for defining sets of content area objectives and for constructing items to measure those objectives. NAEP also has a solid track record in planning and conducting coordinated, standardized national data collections using complex matrix-sampled designs.

Significant economies could be realized by tying the item generation function to an ongoing data collection. Tryout and revision are essential to item development, and for the system envisioned, item calibration is also required. NAEP's matrix-sampled data collections would provide an ideal vehicle for administering new items together with existing items, permitting both tryout and calibration. Analysis of examinee performance on the old and the new items together would permit detection of technically flawed or other anomalously functioning items, and parameters of new items could be estimated by linking them to scales defined by items calibrated previously.

NAEP would be seriously weakened if independent efforts at domain definition and item generation were initiated. Unless the NAEP item pools were linked to that independent system, a longitudinal data collection extending back nearly 20 years would be interrupted.

Domain Descriptions in Different Content Areas

Dimensions of curriculum organization. To be maximally useful, domain descriptions should reflect the natural structures of the curriculum in different content areas. They should permit the representation of content at various levels of aggregation, from a few broad, general indicators of overall achievement down to the smallest separable content elements. They should also reflect the traditional organization of learning outcomes according to skills. The well known Taxonomy by Bloom, et al. (1956) of knowledge, comprehension, application, analysis, synthesis, and evaluation provides one such skill classification, and the distinction it supports between lower-order and higher-order skills is salient in current testing and curriculum policy debates. Perhaps more useful is the related dimension of referent generality (Haertel, 1984; Snow, 1980). Abilities narrow in referent generality can be acquired in a short time and are applicable in a narrow, well-defined range of contexts. At the high end of this continuum are generalized learning abilities, more nearly resembling general aptitudes, which may be the product of years of experience with many types of content in a wide range of learning situations. Yet another dimension along which curricula are organized is that of time, or instructional sequence. In some content areas, this instructional dimension corresponds well to that of complexity or skill level, but in other areas sequencing is more arbitrary (and more variable), and connections to skill level and to referent generality are weaker.

It is not clear how domains could be simultaneously organized according to all these crosscutting dimensions. The general patterns of organization should probably be hierarchical, with narrow objectives grouped together at successively higher levels of generality, but optimum organizations would differ somewhat from one content area to another. In specifying the domains, there would be no need, nor would it be desirable, to tie objectives to age or grade levels, or to specify instructional sequences. The timing of instruction would clearly be relevant to decisions about what content to test at particular grade levels, but the definition of the knowledge base itself would be prior to such considerations.

Advantages of IRT formulation. The problem of domain organization is one which NAEP has addressed with only limited success, primarily due to an absence of any statistical model relating items to more fundamental dimensions of examinee performance. Within NAEP, the organization of items according to objectives is just a taxonomic convenience. Even within objectives, the possibility is acknowledged that separate items may each be of interest in their own right. In the 1978 NAEP mathematics assessment, for example, exercises are tied to objectives, which in turn are classified into content areas (numbers and numeration; variables and relationships; shape, size, and position; etc.) as well as process categories (knowledge, skill, understanding, application). The four process categories are divided into subtopics (recall facts, translate statements, routine problems, etc.) and some subtopics are divided still further. Independent of these hierarchical schemes, other classifications such as "consumer math" or "hand calculator" are imposed at the exercise level. Item difficulties (p values) and average p values are reported for individual exercises and for sets of exercises organized according to "report topics," which include several of these multiple categories, but the overall reporting scheme lacks coherence.

Without some theoretical basis for relating the separate abilities measured by different items, little more can be done--if there are potentially as many abilities to be measured as there are items, then data reduction is problematical. For the domains proposed in this paper, IRT models will be used to implement a conceptual framework permitting better methods of score definition and reporting. Items are conceived in IRT as multiple indicators of underlying latent traits, and item responses are used to estimate some smaller number of trait scores or score distributions.

Objectives as units of domain organization. It is proposed here that the fundamental units for organizing items would be objectives, and these would be of two kinds: simple and composite. Each item would represent exactly one simple objective. Although the meaning of scores on these objectives would be communicated by reference to individual items, no item would be considered important <u>except</u> as a measure of its simple objective. A basic principle of domain organization would require that items measuring each simple objective be homogeneous with respect to content, skill, and format (although they might represent a range of difficulties). This would assure that at the lowest level of domain organization, items defined homogeneous learning outcomes. Composite objectives would be defined as aggregations of simple objectives, of other composite objectives, or both.

The sole purpose for hierarchical grouping of simple objectives into composites would be to define meaningful and significant outcomes spanning broader ranges of content or skill. There would be no necessity to include all objectives within a single hierarchical arrangement, although for most domains the highest level composite would probably be a single, general score for the entire domain, and it would seem desirable to represent all simple objectives in such a score.

It is important to distinguish levels of organization in such a hierarchy from levels of skill. An aggregation of simple objectives testing factual knowledge would still represent only factual knowledge. Higher-order skills would be represented by their own distinct objectives, beginning at the lowest level of domain organization.

<u>Importance of weights in defining composite objectives</u>. The definition and reporting of simple objectives would be a largely technical process, but the definition of composite objectives would require assigning relative weights to their constituent parts, and this process would necessarily involve value judgments. In the scheme proposed, this problem would be made explicit and would be addressed directly. It should not be difficult to arrive at consensus, because composite scores defined according to a range of reasonable weighting schemes would probably be highly intercorrelated. Moreover, dissenting users could always in principle define some alternative weighted composite for their own use.

Weighting questions would become most problematical at higher levels of aggregation, where component scores were more disparate. In U. S. history, for example, some simple objectives might call for factual recall concerning different events or periods, and others might call for critical analyses of these same events or periods. Aggregation of recall objectives across periods, and of analysis objectives across periods, might each be relatively simple. It could prove more difficult to reach consensus on the proper weighting of recall versus analysis in defining a single, overall U. S. History score.

Aggregation according to dimensions other than the major hierarchical classification scheme could be managed in the same way. Continuing the U. S. History example, the primary classification might aggregate each of factual recall and critical analysis across historical periods, as just described. If a score were desired for the Civil War, however, recall and analysis objectives for that period could be combined instead. Explicit assignment of relative weights for these objectives would again be required.

Relationships among objectives. Thus far, objectives within domains have been discussed as if each could be mastered independent of any others. It may sometimes happen, however, that a high level of performance on one objective necessarily implies a high level of performance on others, or perhaps even that poor performance on one objective implies poor performance on others. (This latter inference is more problematical, because poor performance could result from causes other than an absence of skill.) When such logical entailments exist among objectives, it may not always be necessary to test those objectives on which performance can be inferred. In particular, economies might result from the use of complex items that required the exercise of numerous component skills. Such complex items would be organized into their own objectives and described by content, skill and format specifications like any others, but would not necessarily be included in any composite objectives.

Implementation

If a system of achievement domain specifications and item pools is to serve the functions envisioned, there must be broad consensus that it is comprehensive and technically sound. Such credibility can only be attained by careful, systematic implementation. The interests of the Department of Education, of the States, of major teaching subject area organizations (NCTE, NCTM, NSTA, etc.), and of other constituencies must all be represented. Moreover, the different knowledge bases and varieties of expertise represented by disciplinary scholars, curriculum specialists, policymakers, and psychometricians must all be brought to bear in a coordinated fashion if the effort is to succeed.

Attention will also be required to the context in which the proposed system is implemented. If it is to be useful for guiding educational policy, an assessment program must be integrated with the measurement or monitoring of educational inputs and of other kinds of outcome measures, as part of a comprehensive system of national educational indicators (Smith, 1984).

The issues sketched below will require further consideration, but more detailed treatments are beyond the scope of this paper, or would be premature.

How comprehensive should outcome domains be? NAEP has traditionally sought to be eclectic in its inclusion of educational objectives, but in fact has often failed to adequately represent higher-order, critical and analytical skills. Broad representation of cognitive learning outcomes at all levels is essential in a system that may be used to influence curriculum coverage across the Nation, but the degree of comprehensiveness required will depend upon the uses envisioned. The cost of the system, in respondent time and in dollars, will be a function of its scope; if too much is attempted, it will not be done well. Tradeoffs between scope and frequency of assessment must also be considered. It has already been stated that State-level comparisons would probably be based upon some subset of core objectives, and these might be assessed annually or at least biennially. Additional areas might be assessed less frequently, in rotation, following a plan like that of the current NAEP, but with somewhat greater consistency.

Must items be developed for all objectives? There is a risk that if domain specification is unduly burdened by the necessity of providing content, skill, and format specifications for items associated with each objective, then objectives difficult to assess will be slighted. An approach that might obviate this concern would be to develop domain specifications that were broader than the initial set of item pools. These comprehensive domain specifications would provide a framework for continued item development, and over time, item pools could be provided for more and more of the objectives included. It would be explicit at any time what objectives were and were not assessed. Even under this approach, however, it would be helpful if measurement specialists as well as disciplinary scholars and curriculum specialists reviewed domain specifications.

Should affective outcomes be included? The charge of NAEP has been to assess the knowledge, skills, understandings, and attitudes of young Americans, and all NAEP assessments have included items measuring attitudes and other affective learning outcomes. Incorporation of affective outcomes in domain structures like those discussed is clearly problematical. At this point, the assessment of attitudes and values is simply set forth as an issue requiring further consideration.

What IRT models should be used? Should these models be applied at the level of simple objectives, or of composite objectives? Dozens of IRT models useful for different purposes have appeared in the psychometric literature of recent years, but the major contenders for applications like those envisioned here are the one-parameter logistic (Rasch) model, the two-parameter logistic or normal ogive models, and the three-parameter logistic or normal ogive models. There is no need to review the extensive debate over the relative merits of these models, or to discuss alternative estimation procedures, but decisions on these matters will have to be reached in order to implement the system proposed.

An assumption common to all of these models is that items calibrated together all measure the same ability, that all can be appropriately referenced to the same latent skill continuum. Statistical tests of unidimensionality are available for some of these models (e.g., Gustafsson, 1980; Molenaar, 1983), but by and large these are not entirely satisfactory. In the system proposed, the unidimensionality assumption might be taken to imply that scaling and calibration should only be carried out at the level of the simple objectives. Scores for composite objectives would then formed by taking weighted averages of the scores on their constituent objectives.

This approach would indeed assure the psychometric purity of each scale, but the alternative of scaling fairly homogeneous composite objectives directly could be considerably less expensive. There are practical limits to the minimum number of items that can be calibrated together, below which accuracy becomes intolerably poor and some estimation procedures fail altogether. These limits vary from one IRT model to another, and among estimation methods for the same model, but it is likely that the minimums would exceed the number of items otherwise required for at least some simple objectives. It would seem wasteful to construct and administer extra items solely to permit calibration at the level of simple objectives.

The alternative of calibrating composite objectives directly would be more cost effective, and could improve accuracy, but at the same time could increase violations of model assumptions, introducing bias. Such tradeoffs between bias and precision are common in statistical modeling, and compromise is required. Different IRT models vary in their degree of tolerance for violations of the unidimensionality assumption, but in general, the two- and three-parameter models have proven quite robust.

If simple objectives narrow in referent generality (and therefore most sensitive to instruction over the short term) are aggregated, model violations might be minimized by appropriate timing of the data collection. Suppose, for example, that a set of factual objectives covering different periods of U. S. History were to be scaled together. If data were collected mid-year, then items testing content not yet covered might appear too difficult relative to items testing material covered earlier. Data might be collected at the end of the year, or students currently enrolled in U. S. History might be excluded from the calibration sample; neither of these alternatives would be entirely satisfactory.

<u>What models other than unidimensional IRT models might be considered?</u> It would surely be unwise in a large-scale application like the one proposed to rely upon statistical methods still considered experimental. The IRT models described above have been in wide use for many years, are well understood, are widely accepted, and if intelligently applied, should serve admirably for the proposed assessment system. Nonetheless, some promising new models of quite different kinds might also be considered for limited use. Foremost among these would be multidimensional IRT models, which are in effect factor models for dichotomized variables, and latent class models for item response data.

Multidimensional IRT models posit two or more latent dimensions, and include item parameters relating correct response probabilities to each of these dimensions. Such models would find natural application in scaling composite objectives that were truly multidimensional, but might also be used for occasional simple objectives for which items assess different mixtures of two or more proficiency dimensions despite homogeneous specification of content, skill, and format.

Latent class models differ from latent trait models in positing underlying variables that are categorical rather than continuous. Multidimensionality is easily accomodated using such models, and interactions among dimensions can be handled quite flexibly. Each dimension, however, is represented by a mastery-nonmastery dichotomy, or at most by a small number of skill levels. Such models have shown considerable promise for describing achievement test performance, and merit further investigation (Haertel, 1986).

<u>Standards and criteria</u>. A useful distinction can be drawn between (1) the scales, or skill dimensions, defining curricular attainments and (2) the levels or degrees of proficiency along those dimensions that are designated

acceptable. The former are criteria, and the latter are standards.

For purposes of standard setting, there are at least three reasons why scales like those proposed, defined by reference to item pools, are superior to scales defined with reference to intact tests. First, such scales have greater permanence. Experience with examinees at different performance levels can accrue over time, making empirical studies to establish the correlates of different performance levels more cost effective. Second, score scale points can be illustrated by reference to specific calibrated items ("A score of 250 means an examinee has an 80 percent chance of answering one of these items correctly. A score of 300 means an 80 percent chance of answering these correctly. . . ."). This permits persons charged with standard setting to better understand just what different scale points mean in behavioral terms. Third, descriptions of the meaning of different scale points can be developed, based upon the common characteristics of items located at those levels, as was done for the 1984-85 NAEP reading scale (Beaton, 1986; Reading Report Card, 1985).

Planning will be required to capitalize on the potential of domains and item pools for standard setting. Norm-referenced comparisons of schools, districts, or other aggregates to one another and to their own previous performance can be reported just as they are in most existing State testing programs. In addition, however, it will be possible to establish more defensible absolute targets or levels of acceptable performance, based on descriptions of the actual proficiencies represented and on evidence of their behavioral correlates. It may be useful to establish a series of definite proficiency levels, like the five levels of the 1984-85 NAEP reading scale, so that reporting can include targets relevant to a range of grade levels, pupil proficiency levels, and so forth.

Notes

[1]The term "item" in this paper refers to the same broad range of assessment tasks referred to in NAEP as "exercises."

References

Beaton, A. E. (1986, April). The NAEP reading scale. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

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

Educational Testing Service. (1985). The reading report card, Progress toward excellence in our schools, Trends in reading over four assessments, 1971-1984. Princeton, NJ: Author.

Gustafsson, J-E. (1980). Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 33, 205-233.

Haertel, E. H. (1984). Construct validity and criterion-referenced testing. Review of Educational Research, 55, 23-46.

Haertel, E. H. (1986). Using restricted latent class models to map the skill structure of achievement items. Manuscript submitted for publication.

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. (1974). Anchor test study- The equating and norming of selected reading achievement tests (grades 4, 5, and 6). Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education.

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. Psychometrika, 48, 49-72.

Reading Report Card (NAEP Report No. 15-R-01). (1985). Princeton, NJ: Educational Testing Service.

Smith, M. S. (1984, November). A framework for the development of national educational indicators. In Council of Chief State School Officers, Education evaluation and assessment in the United States. Position paper and recommendations for action. Washington, DC: Author.

Snow, R. E. (1980). Aptitude and achievement. In W. B. Schrader (Ed.), Measuring achievement: Progress over a decade. New Directions for Testing and Measurement (No. 5). San Francisco: Jossey-Bass. (Proceedings of the 1979 ETS Invitational Conference)