DOCUMENT RESUME

ED 342 821 TM 018 142

AUTHOR Chelimsky, Eleanor
TITLE National Assessment Governing Board (NAGB) Achievement Levels. Interim Letter Report.
INSTITUTION General Accounting Office, Washington, DC. Program Evaluation and Methodology Div.
REPORT NO GAO/PEMD-92-22R
PUB DATE 11 Mar 92
NOTE 20p.; Letter to the Committee on Education and Labor and the Subcommittee on Elementary, Secondary, and Vocational Education of the House of Representatives.
PUB TYPE Legal/Legislative/Regulatory Materials (090) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Academic Standards; Data Analysis; Government Role; Letters (Correspondence); National Programs; Program Evaluation; *Reliability; *Research Methodology; Research Reports; Testing Programs; *Validity
IDENTIFIERS *National Assessment Governing Board; National Assessment of Educational Progress; National Center for Education Statistics; Standard Setting

ABSTRACT

This letter is an interim response to the October 7, 1991 request from the Committee on Education and Labor and the Subcommittee on Elementary, Secondary, and Vocational Education of the House of Representatives asking for a review of the National Assessment Governing Board (NAGB) achievement levels for the National Assessment of Educational Progress (NAEP). It summarizes findings and conclusions to date. The General Accounting Office (GAO) reviewed the record of the development and results of the NAGB's level-setting approach from early 1989 to the present and interviewed government and testing contractor (Educational Testing Service) staff. In examining the achievement levels, the GAO found problems of procedure, reliability, validity, and reporting. Problems seem sufficiently serious to preclude continuing the application of the achievement levels. It is contended that too much is uncertain to support the NAGB's decision to organize the reporting and analysis of the 1992 NAEP results around achievement levels. Even with continued improvement in the item judgment method of setting standards, the approach does not seem suitable. The NAGB apparently conducted only a cursory review of alternative methods for setting standards. Recommendations are made for realigning the functions of the NAGB and the National Center for Education Statistics or for strengthening the capacity of the NAGB to make sound technical decisions. Three enclosures provide additional information about the NAEP and the achievement levels. (SLD)

GAO

United States
General Accounting Office
Washington, D.C. 20548

Program Evaluation and
Methodology Division

March 11, 1992

The Honorable William D. Ford
Chairman, Committee on Education and Labor
House of Representatives

The Honorable Dale E. Kildee
Chairman, Subcommittee on Elementary,
Secondary, and Vocational Education
House of Representatives

On September 30, 1991, the National Assessment Governing Board (NAGB) released a report that interpreted U.S. students' achievement in mathematics on the 1990 National Assessment of Educational Progress (NAEP) in terms of a set of performance standards. NAEP, a nationwide test funded by the Department of Education and administered by the National Center for Education Statistics (NCES) under NAGB's direction, had measured student achievement in basic subjects since 1969 without reference to specific goals or standards. However, the 1988 legislation that created NAGB also made it responsible for "identifying appropriate achievement goals." NAGB designed and implemented an approach to define basic, proficient, and advanced levels of achievement and to express each level in terms of a score on the 1990 mathematics assessment. It reported proportions of students who reached each level even though its evaluation consultants and others had expressed concerns about the appropriateness of the NAGB approach and about the technical quality of the 1990 results.

This letter is an interim response to your October 7, 1991, request for a review of the NAGB achievement levels and related matters, summarizing our findings and conclusions at this point in our work. Our full report will be completed later this year.

Your letter noted the wide interest in issues of how to set standards and measure progress toward them, interest that is continuing in view of the recommendations of the National Council on Educational Standards and Testing (NCEST) for more of both such efforts. It is important that pioneering efforts such as NAGB's be fully examined and their strengths and limitations explicitly identified so as to provide the soundest possible guidance to future definitions and applications of standards.

At your request, we are examining how NAGB set the mathematics achievement levels used in reporting the 1990 results, including in our assessment a review of the validity of criticisms leveled by NAGB's evaluation consultants. You also asked for a broader analysis of the Board's resources and procedures for technical quality control in its work. The specific questions we are addressing in response to your request are:

-- Were the NAGB 1990 achievement level-setting exercise and its products flawed?

-- Is the NAGB approach for setting achievement levels or standards suitable for use with the NAEP test?

-- Are NAGB's resources and procedures sufficient to ensure that work done at its direction and the products that result are technically sound?

GAO'S REVIEW TO DATE

We reviewed the record of the development and results of NAGB's 1990 levels-setting approach from early 1989 to the present and interviewed NAGB staff, the principal NAGB consultant for the levels activity, and technical staff at NCES and of its testing contractor, the Educational Testing Service (ETS). We reviewed NAEP technical procedures and consulted the literature on setting performance standards. We looked for evidence that NAGB's achievement levels findings are consistent with its definitions and with related data. We also examined the NAGB evaluation team report, the report on the first phase of the levels project by the NCES Technical Review Panel for Studies of the Validity of NAEP, and the report of the National Academy of Education Panel on the evaluation of the NAEP Trial State Assessment. We spoke with the authors of these reports and with other experts. Finally, we are conducting but have not yet completed an initial review of NAGB action in two additional decision areas.

THE 1990 ACHIEVEMENT LEVELS EXERCISE

We evaluated the 1990 levels exercise--which involved expert judges' examining NAEP test items in mathematics to set three levels of achievement--against standards for reliability, validity, and reporting of technical data. We looked to see whether NAGB's procedures were similar to those commonly followed in connection with item-judgment methods, and we asked whether the results have been useful. We found problems of procedure, of reliability, of validity, and of reporting. We examined plans for further work of this sort in light of these problems and

concluded that commitments to the further use of levels now being set in similar fashion seem premature. These findings concerning our first evaluation question are discussed in the paragraphs that follow.

Procedures and Results

With respect to procedure, we found that NAGB asked members of item judgment panels to conceive of three subgroups of students, each representing a newly formulated level of performance that was only generically defined, and to estimate the proportion of students just qualified for each subgroup that would answer each question on the 1990 NAEP mathematics test correctly. (The levels definitions are shown in enclosure I.) This approach extended the item judgment methodology beyond its usual setting. Item judgment procedures have been most often used to determine a single standard of performance that qualifies examinees for certification (for example, as a high school graduate or entry-level professional). According to the literature, item judgment methods work best when judges share a clear conception of the just-qualified group.

As NAGB's technical report on the project recognized, the generic definitions of the three levels did not provide a clear conception of the degree of performance that is meant by each term (basic, proficient, advanced). Thus, they did not provide a clear standard for judging the proportion of students at each level that would be likely to answer each item correctly. Differences in interpretation may well be responsible for weaknesses we found in the reliability of the results of the judgment process.

An item judgment method can be said to be reliable when there is evidence that, if repeated, it would produce the same results as before: that is, that a second panel of judges selected on the same criteria as the first panel and judging the same items would come up with the same estimate of how students will perform on a test. Reliability is decreased when individual panelists are inconsistent in their judgments and when judgments vary substantially from panelist to panelist. NAGB's technical report comments that there was "substantial and troublesome" variability in estimates of the basic level for each grade across the four panels whose judgments form the basis for the 1990 levels analysis. Different panels, in other words, produced differing estimates--an indication that reliability of the results may warrant further examination. The reliability of the estimate of the percentage of students achieving the advanced level is

also uncertain, because the level falls at the extreme high end of the distribution of actual NAEP scores where there are too few data to sustain a confident estimate.

We found that the empirical validity of the levels results--whether the levels measure what they claim to measure--has not been demonstrated. Item judgment results, even if reliable, are not necessarily empirically valid. To know whether a score derived from the judgment process validly distinguishes examinees who have met the standard from those who have not requires that the results be compared to other sources of information and adjusted if necessary. This is especially so when a standard is predictive of performance--that is, when it is claimed, as NAGB claims in its levels definitions, that someone who meets the standard will perform successfully and that someone who scores below the standard will not. We find that NAGB has taken no action to assess the empirical validity of the levels results, and hence that their predictive accuracy is as yet unsupported.

To conduct an initial check of the validity of the levels that NAGB set, we compared the levels results to other indicators of mathematics achievement and found warning signs. For example, even in fourth-grade classes identified by their teachers as having high ability, fewer than 5 percent of the students achieved at the advanced level; in the same classes, more than 10 percent did not reach the basic level (had not even partially mastered skills fundamental to proficient work). Data from an international test suggest that the advanced level identified for fourth and eighth grade students exceeds even world class standards. (Further information on these comparisons is presented in enclosure II.)

Finally, we found that NAGB did not disclose the limitations of the levels data when it published the 1990 mathematics levels results. The report released September 30, 1991, did not caution readers concerning the reliability of the data nor note that validity had not been established. NAGB's technical report with more of such information was not available until late November, 1991. Documentation of the quality of the data was important, we believe, in view of the problems listed above. NCES officials told us that the NAGB standard-setting results, which were published under NAGB's independent authority, probably would not have passed NCES's pre-publication statistical quality review.

Whether the levels data are useful remains to be seen. The National Academy of Education has conducted interviews with consumers of NAEP data, but the results are not yet

available. NAGB has surveyed users and analyzed press coverage of the achievement levels results, and has found widespread interest in the levels but also some errors in interpretation of the results. The National Educational Goals Panel, which planned to be a major user of the new standards, adopted the proficient level as an indicator of competency in mathematics in its first panel report, but did not adopt the basic or advanced levels based on concerns over their soundness.

The Continued Application of the Policy

Given the problems we found with NAGB's 1990 procedures and their results, we are concerned about plans going forward for more such work. Procedures to apply the achievement levels approach to mathematics, reading and writing for the 1992 assessment are under way. The Department of Education has awarded a $1.34 million contract to an experienced organization, American College Testing (ACT), whose proposal provides for many improvements over the 1990 procedures including the development of subject-specific descriptions of the three achievement levels to guide the item judgment process and a thorough examination of the reliability of the judgment results. However, ACT has not been directed to assess empirical validity nor will that be possible in the short time available in the current schedule for analyzing and reporting 1992 NAEP results in these three subjects according to the levels.

The NAGB approach for setting performance levels is being tried for the first time with NAEP reading and writing tests, which involve question formats and scoring and scaling issues not encountered in the mathematics assessment. Given unresolved questions about NAGB's approach and new technical issues posed in the reading and writing assessments, there is no assurance that the judgment procedure will produce technically acceptable and usable results. The results are to be published as part of the NAEP report, and to form the basis for the tabular displays and interpretive text describing overall student achievement at the national and state levels and for subgroups of students. As part of the NAEP report, the levels will have to meet NCES statistical quality standards. If they do not, the levels data cannot be approved for release. Problems in the technical quality of the levels results could delay the issuance of the 1992 NAEP report substantially.

Too much is uncertain, we believe, to support NAGB's decision to organize the reporting and analysis of 1992 NAEP results around achievement levels that may or may not

be found to be workable. Such a decision risks delay and additional cost, and it also risks approval of levels results before they have been thoroughly examined in order to meet a publication schedule. We consider these risks unnecessary because there seems no compelling reason to publish the levels results at the same time as the purely descriptive portions of the NAEP: that is, the report of average scores and of the distribution of scores. Much might be gained, at a modest additional cost, by taking the time to conduct a full evaluation and validity review before releasing the levels data or other interpretive materials.

IS THE NAGB METHOD FOR SETTING STANDARDS GENERALLY SUITABLE FOR USE WITH NAEP?

Even if NAGB and contractors make continued improvements in the item judgment method of setting standards, we found reasons on a variety of criteria to question whether this is a suitable general approach.

First, the technical requirements of the approach may conflict with NAGB's charge to develop goals through a broad consensus process. The most reliable results in item judgments come from panels composed of experts who have a common and well-informed frame of reference for their decisions. (NAGB progressively narrowed the participants in the 1990 process, when it was found that non-educators gave somewhat different judgments than did teachers.) Thus, the judgment process does not yield broad consensus on the appropriateness of the results--on whether the advanced level, for example, is reasonable in terms of student exposure and consistent with other indicators of advanced performance. NAGB's approach does not provide for additional consensus processes prior to the judgment process (to identify content standards against which item judgments can be made). The 1992 plan does provide for public comment on the results of the judgment panels' work (that is, on the number and descriptive paragraphs that pertain to each grade and level of achievement). As we understand it, information showing the test score corresponding to each level and the proportion of students achieving the level will not be public at that time.

Second, the three-level approach chosen by NAGB--combined with pressures to move in the direction of framing test coverage in terms of curricular ideals rather than in terms of current practice--may pose technical conflicts with NAEP's purpose and design. NAEP tests have until recently been designed to describe the overall distribution of achievement with the greatest accuracy in

the middle range. (Accuracy is in part a function of the number of test items; to get the greatest accuracy for the greatest number of students, NAEP has had relatively more items in the moderate range of difficulty than at the extremes.) Thus, NAEP tests have not been designed to discriminate accurately at the far upper end of the distribution as the advanced NAGB level would seem to require. As NAGB recognizes, going further with the three-levels policy may require changes in NAEP's design, such as expanding the number of harder items to improve measurement at the upper extreme. If that led to fewer items for the average or below-average student, such changes could decrease NAEP's overall reliability.

Third, NAGB's approach has a general problem of utility: although new information is provided about how many students reach the three standards of overall mathematics proficiency, the levels do not apply to the detailed mathematics sub-areas such as algebra or numbers and operations. Only the composite scale offers enough data to sustain a judgment about the advanced level. Thus, unless the number of test questions in each component skill or content area is expanded substantially, reporting on overall student performance by level cannot help educators identify specific areas of weakness in student performance. Furthermore, if the component subjects do not form a composite scale (which is increasingly likely as NAEP addresses new types of skills areas and new forms of assessment), NAGB's approach may not apply.

The overall usefulness and suitability of NAGB's approach may need reconsideration in light of developments since it was adopted in 1990. NAGB's approach derives performance levels from NAEP test items which traditionally have reflected what is currently being taught, according to general definitions of competence. However, content and performance standards are increasingly likely to be set by subject area groups or by states in the near future. Such standards may well include skills and content not currently covered in the majority of classrooms, may address component skills and content areas separately, and may not conceive of three levels of achievement on a composite scale. Applying such standards to NAEP will pose many new issues, and may require using a combination of approaches.

We conclude that NAGB's approach has significant limitations and that its future applicability may be limited. Continued pursuit of a policy and methodology that have demonstrable limitations seems less useful than starting now to consider and test alternatives in light of changes in educational standards and assessment since the

policy was adopted.

THE TECHNICAL QUALITY OF NAGB's DECISIONS

To answer the third evaluation question, we reviewed the process through which NAGB formulated, applied, adjusted, and acted upon the results of its achievement levels policy. We identified key decision points and the technical resources and information that were used at each point, and we evaluated whether that information was sufficient to support the decision. Our review found that NAGB undertook an innovative and demanding technical procedure with the slenderest of technical resources: chiefly, one staff psychometrician and one part-time consultant. Experts in NCES and ETS and outside of government assisted with particulars, but concentrated outside assistance was available only after the 1990 project was well under way. We found evidence that key decisions were made before their technical soundness had been adequately examined and that technical quality criteria were not clearly defined or applied.

We conclude from staff papers, NAGB meeting transcripts, and committee records that NAGB conducted only a cursory review of alternative methods for setting standards and did not analyze the limitations of the item-judgment method as the sole basis for standard-setting. These matters were amply discussed in technical literature available at the time.

While NAGB devoted considerable thoughtful discussion to the desirability of having multiple standards one of which would be reached only by the most advanced students, the technical implications of this choice for use with item judgments on NAEP data were not systematically examined before the policy was instituted. From what we can determine, NAGB simply assumed that having judges make three ratings per item guided by generic definitions of the three levels would not be a problem, although the technical literature suggested otherwise.

Although an early concept paper and the May, 1990 levels policy were circulated for comment, NAGB did not formally develop a technical design for the initial levels judgment process and obtain a technical critique before going ahead. (NAGB did get reviews of the materials developed to orient judges, but not on the overall plan.) The redesign for four regional meetings, developed after the initial design proved problematic, was also only sketchily developed and reviewed in advance. The redesign noted that reviewers had raised technical questions about the basic approach but did not respond to those questions.

We found that NAGB decided in 1991 to go ahead with further application of the unproven methods in 1992 (and solicited the contract proposals) before the reliability of the levels results had been fully evaluated and in the absence of any information to confirm their validity. Bidders could have been asked, but were not, to address questions about the overall policy raised by the evaluation team and others.

NAGB did get good advice. Most of the methodological and procedural issues that proved problematic were raised during the period in which the policy was under consideration, many of them by NAGB members. Much good advice was followed: NAGB instituted many improvements in its item-judgment procedure as problems with the initial conception became evident. However, NAGB did not respond to fundamental questions about the approach itself or to calls for the consideration of alternatives. Although NAGB initially emphasized the provisional nature of its approach, it has not reconsidered its approach nor has it provided evidence that the fundamental criticisms of it are in error.

There are established standards for the development, evaluation and reporting of the results of item-judgment procedures. These standards were only partially met by NAGB's reports. It is not clear whether NAGB considered these standards fully applicable to what it considered to be a trial venture. NAGB has adopted technical standards for item development and review and for data collection, but it does not have standards for reporting statistical data.

Finally, the record indicates that while some NAGB members and staff understood the methodological issues involved in this exercise, some did not or thought they had not been given sufficient information about the proposed approach at the time of the policy decision. As one member stated in the May 1990 meeting, anyone who understood the approach should "be classified as Advanced."

It is true that the levels example may be unique. Setting performance standards was new for NAEP, so there was little direct experience to guide NAGB's deliberations. NAGB elected to conduct the 1990 exercise with its own limited technical resources. Users such as the National Education Goals Panel were eager to get results. The technical deficiencies noted in this letter may reflect these particular circumstances. To determine whether this is the case, we are reviewing NAGB decisions in additional technical areas. Our review will concentrate on the decision procedures, on the technical resources applied at

key decision points, and on technical quality control mechanisms and their application.

CLARIFICATION OF THE ROLES OF NAGB AND NCES

In search of explanations for the problems we observed, we examined NAGB's organization structure and statutory responsibilities and those of NCES with respect to NAEP to determine whether a lack of clarity in these basic arrangements may have contributed to the difficulties observed in the levels-setting example. Where we found a lack of clarity, we have considered options that might clarify the roles of these two units in a manner that could strengthen the overall operation of NAEP.

The present structure of two units, each with a unique strength, has the potential for continuing difficulty. (This structure is shown in enclosure III.) One unit, NCES, is staffed by full-time technical experts and has access to others through its own advisory committees, through ETS as the technical contractor for NAEP, and through the technical advisory body that ETS is required to consult. The other, NAGB, is a lay body composed of members of key constituencies (state and local officials and educators, citizens, and two experts in measurement) that meets several times a year with committee activities between meetings; it is assisted by a staff of six, of whom two are technical experts.

NCES implements a highly technical project, NAEP, with advice from NAGB. NCES is responsible for ensuring that the Assessment is fair and accurate. The Commissioner of Education Statistics, who heads NCES, is the guardian of the quality of statistical data produced under his supervision (including NAEP data). NCES also conducts review and validation studies of NAEP and solicits public comment on its usefulness.

NAGB is responsible for formulating policy guidelines for NAEP. It also has a number of specific responsibilities. Some of NAGB's specific responsibilities reflect the kinds of review and broad direction well suited to a broadly representative, part-time body: for example, ensuring that items are free from bias and that goal statements for each learning area are developed through a consensus approach, or by creating guidelines for reporting. Other responsibilities appear to be highly technical, such as "designing the methodology of the assessment." While a part-time body composed primarily of non-experts might usefully suggest broad changes in design (such as the inclusion of private schools or of out-of-school

adolescents in the NAEP sample), it is difficult to envision how such a body could effectively develop and implement technical operations within its own resources.

The law does not limit NAGB to policy direction: it seems to suggest that it engage in technical design as well. This creates the possibility, illustrated in the levels case, that NAGB could give NCES technical direction that, if faithfully implemented, would produce results that NCES, in its role as guardian of statistical quality or of fairness and accuracy, could not approve for release.

There appear to be two main ways to clarify this situation: either realign functions to concentrate each unit's efforts on the activities for which its current composition and resources are best suited, or strengthen NAGB's capacity to make sound technical decisions. The first alternative may be more feasible than the second.

Realigning the Functions Assigned to NAGB and NCES

NAGB's current strength lies chiefly in the breadth of perspective and links to NAEP consumers and providers (that is, those whose classes are tested) that its members provide. These strengths make NAGB a strong and credible source of broad policy guidance on such issues as

-- the scope and coverage of NAEP (for example, subject areas and student populations to be included in the assessment),

-- information, analysis and reporting (what consumers want to know, what forms and methods of presentation reach them most effectively), and

-- uses of the assessment (comparisons, linkages, cooperative ventures).

NAGB's broad membership can also be an asset in

-- evaluating proposed technical changes in NAEP from provider and consumer perspectives,

-- overseeing the operation of the consensus processes through which achievement goals are selected, and

-- deciding issues of item bias or item appropriateness, if questions remain when the consensus process has concluded.

Assigning NAGB the above functions would give NAGB the primary role in identifying what is to be done, but

relieve it of the burden of determining how things are to be done. NAGB's small staff and limited technical resources are adequate to support these guidance and review functions.

NCES's strength lies in its technical resources. It is well suited to propose technical improvements in NAEP for NAGB's review. NCES is also well suited to provide NAGB with the technical information needed before policies are adopted and implemented: to convene experts to explore an issue, develop alternatives, or evaluate the technical feasibility of a given approach. NCES can then develop and implement an approach once it is found to be feasible and appropriate, and report back to NAGB as the venture proceeds.

Strengthening NAGB's Capacity to Make Sound Technical Decisions

If NAGB retains operational responsibility for technical functions, its capacity to make sound technical decisions could be increased by adding to the technical expertise available to it, by adopting procedures to increase the technical soundness of policy decisions, and by adopting and applying technical quality standards and technical quality control mechanisms.

Technical resources could be increased by

-- adding technical experts to NAGB membership,

-- increasing the size of the NAGB technical staff, or

-- providing for NAGB use of the technical resources available through NCES staff and contractors or through its own contract resources.

The quality of NAGB's decisions could be improved by adopting procedures to ensure that the technical implications of a proposed policy decision are fully reviewed and their cost and feasibility examined prior to the institution of the policy. At a minimum, NCES should be asked to conduct such a review.

Product quality could be ensured by

-- requiring NAGB statistical products to undergo NCES statistical quality review,

-- requiring NAGB statistical products to meet NCES statistical standards (with NAGB obtaining review from NCES or elsewhere and describing standards, reviewer

qualifications, and review results in its product), or

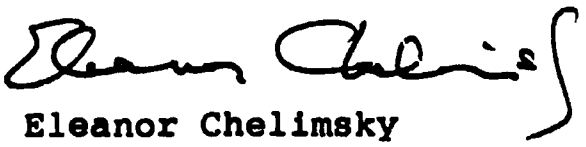
-- requiring NAGB to develop its own standards or adopt existing standards and to describe the standards review and its results in its publications as above.

If NAGB's composition and responsibilities remain as they are, the product quality and decision quality improvements are essential to prevent risk and also to lessen the need for NAGB members to have technical knowledge. However, for its own comfort NAGB might wish to ensure that there is enough expertise within its membership and staff to be able to recognize the quality of the advice it is getting.

We conducted the work described above in Washington, D.C. between September 1991 and March 1992 in accordance with generally accepted government auditing standards.

As you requested, we did not send this interim report to the National Assessment Governing Board and the Department of Education for comment prior to publication. However, we did meet with officials from NAGB and NCES, and we briefed them on our major findings and conclusions. If you have questions or would like additional information, please call me at (202) 275-1854 or Robert L. York, Director of Program Evaluation in Human Service Areas, at (202) 275-5885.

Sincerely yours,



Eleanor Chelimsky
Assistant Comptroller General

Enclosures

NAGB DEFINITIONS OF ACHIEVEMENT LEVELS

Source: National Assessment Governing Board, The Levels of Mathematics Achievement (Washington, D.C., National Assessment Governing Board, 1991), Volume I, p. 5.

"Basic. This level, below proficient, denotes partial mastery of knowledge and skills that are fundamental for proficient work at each grade level--4, 8, and 12. For 12th grade, this is higher than minimum competency skills (which normally are taught in elementary and junior high schools) and covers significant elements of standard high-school work.

"Proficient. This central level represents solid academic performance for each grade tested--4, 8, and 12. It reflects a consensus that students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. At grade 12, the proficient level encompasses a body of subject-matter knowledge and analytical skills, of cultural literacy and insight, that all high school graduates should have for democratic citizenship, responsible adulthood, and productive work.

"Advanced. This higher level signifies superior performance beyond proficient grade-level mastery at grades 4, 8, and 12. For 12th grade, the advanced level shows readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement. As data become available, it may be based in part on international comparisons of academic achievement and may also be related to Advanced Placement and other college placement exams."

COMPARISON OF NAGB LEVELS AND OTHER DATA

COMPARISON USING NATIONAL DATA

The accompanying figure shows the distribution of NAEP scores for fourth grade mathematics from the 1990 assessment for all students and for subgroups of students. Each row on the chart shows the score attained at the 5th percentile (that is, by the lowest 5 percent), the 10th percentile, the 25th, the 50th (midpoint), the 75th, the 90th, and the 95th percentile (the top 5 percent). The scores that set the lower boundary of basic, proficient, and advanced achievement are shown by the vertical lines.

The four horizontal lines or rows show the range of scores for the total population tested and for three subgroups.

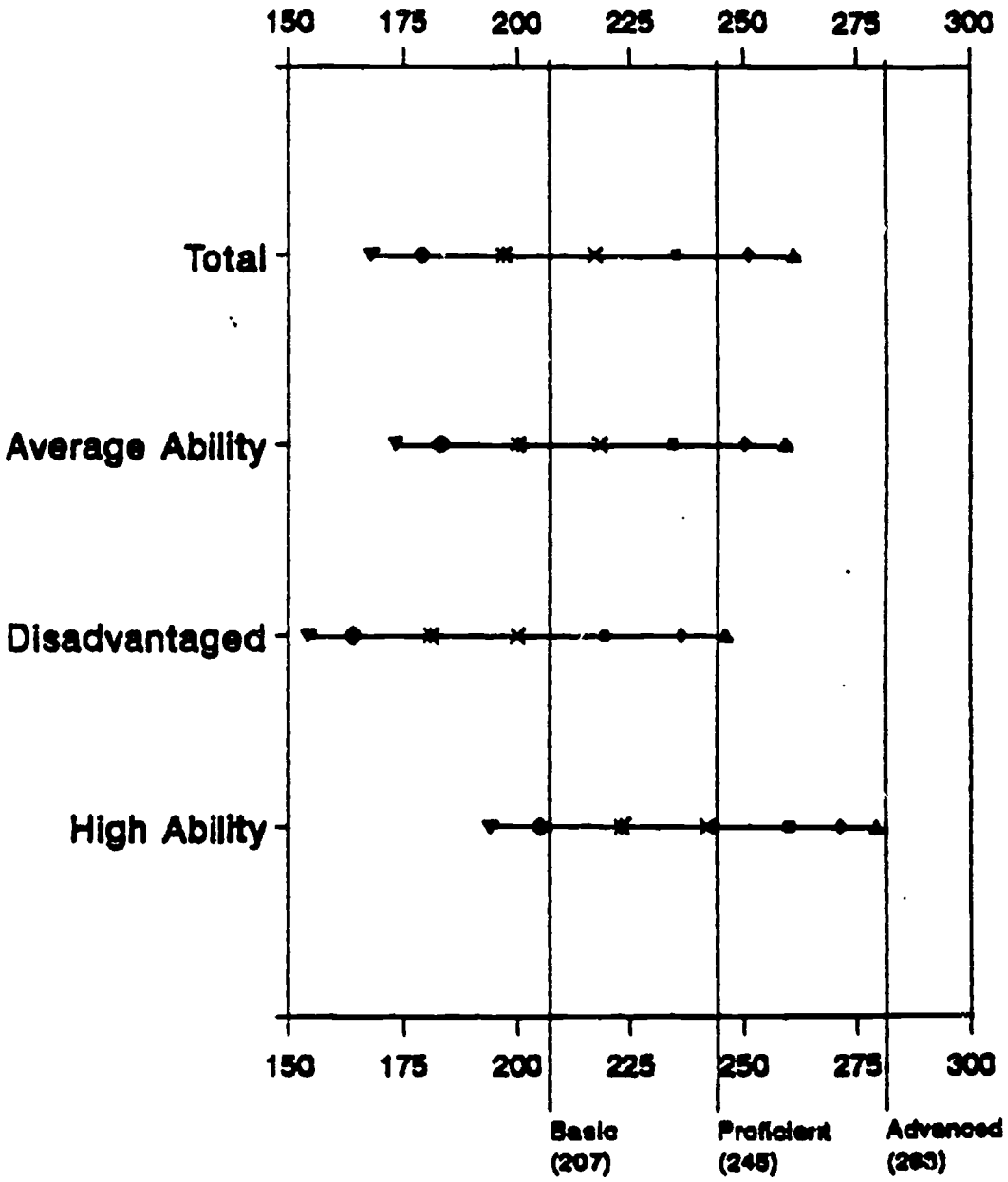
-- Row 2 shows the scores for students in classes that teachers independently identified as being composed of students of average ability. The distribution shown is very similar to the distribution for the total population, suggesting that teachers' judgments of class ability are reasonably accurate.

-- Row 3 shows the scores for students in disadvantaged urban districts. This subgroup is important because NAGB sought to give even the least advantaged students a standard they could strive to achieve.

-- Row 4 shows the scores for students in classes that teachers identified as being composed of students of high ability.

Applying NAGB's levels definitions to these distributions supports the interpretation that more than half of the fourth grade students in disadvantaged urban communities had not achieved even partial mastery of fundamental skills for the grade (basic achievement) and just 5 percent were well prepared to move on to the next grade. More than 10 percent of the students in high-ability classes had not reached the basic level of achievement; fewer than half were proficient and fewer than 5 percent advanced. These interpretations appear to be somewhat extreme and likely to have a discouraging effect for both the very able and the disadvantaged.

NAGB Achievement Levels and NAEP Score Distributions in Fourth Grade Mathematics, 1990



Source: National Center for Education Statistics, The State of Mathematics Achievement: NAEP's 1990 Assessment of the Nation and the Trial Assessment of the States (Washington, DC: U.S. Government Printing Office, June 1991), p. 491.

COMPARISON USING INTERNATIONAL DATA

We examined data on mathematics achievement of 9-year-olds and 13-year-olds from the 1990-91 International Assessment of Educational Progress (IAEP). The IAEP test was similar to NAEP, though the content was adjusted to be reasonably representative of curricula across the various participating nations. To estimate the proportion of students in other nations who might qualify as advanced in NAGB's terms, we identified the score attained by the top 1 percent of U.S. 9-year-olds on the international test (equivalent to the percentile of fourth graders that qualified as advanced in terms of NAEP) and identified the proportion of students from other nations who equaled or exceeded this score. The same procedure was applied to the scores of eighth graders and 13-year-olds. (The IAEP report did not provide sufficient detail to allow similar comparisons at the basic and proficient levels.)

Fewer than 5 percent of the 9-year-old students in any nation tested demonstrated advanced achievement according to this comparison. For 13-year-olds, 10 percent of the students in Taiwan and at least 5 percent of those in China (restricted sample) and Korea met this standard; in no other nation tested did as many as 5 percent meet the advanced threshold. This comparison indicates that the advanced level is extreme even by world class standards.

GOVERNANCE AND ADMINISTRATIVE STRUCTURE FOR THE NATIONAL ASSESSMENT

NAGB membership
- 2 governors or former governors
- 2 state legislators
- 2 chief state school officers
- 1 superintendent of a local educational agency
- 1 member of a state board of education
- 1 member of a local board of education
- 3 classroom teachers
- 1 representative of business or industry
- 2 curriculum specialists
- 2 testing and measurement experts
- 1 nonpublic school administrator or policymaker
- 2 school principals
- 3 representatives of the general public

NAGB responsibilities
- Formulating policy guidelines for NAEP
- Selecting subject areas to be assessed
- Identifying appropriate achievement goals
- Developing assessment objectives
- Developing test specifications
- Designing the methodology of the assessment
- Developing guidelines and standards for analysis plans and for reporting and disseminating results
- Developing standards and procedures for interstate, regional, and national comparisons
- Taking appropriate actions to improve the form and use of NAEP

NAGB also has final authority on the appropriateness of cognitive test items and is directed to ensure that such items are free from bias and that each learning area assessment has goal statements developed through a national consensus approach.

NCES responsibilities regarding NAEP
- Carrying out NAEP with the advice of NAGB
- Ensuring that NAEP provides a fair and accurate presentation of educational achievement, uses representative sampling, reports trends reliably, and includes information on special groups
- Securing an independent evaluation of the Trial State Assessment
- Ensuring the technical quality of the published data
- Conducting reviews and validation studies of the

ENCLOSURE III

ENCLOSURE III

National Assessment and soliciting comment on its conduct and usefulness

(973717)

19 GAO/PEMD-92-22R National Assessment Technical Quality