DOCUMENT RESUME

ED 414 157 SE 060 637

AUTHOR Pashley, Peter J.; Phillips, Gary W.

TITLE Toward World-Class Standards: A Research Study Linking International and National Assessments.

INSTITUTION Educational Testing Service, Princeton, NJ.; Westat, Inc., Rockville, MD.; Educational Testing Service, Princeton, NJ. Center for the Assessment of Educational Progress.

SPONS AGENCY National Science Foundation, Arlington, VA.; Office of Educational Research and Improvement (ED), Washington, DC.; Carnegie Corp. of New York, NY.
REPORT NO ETS-24-CAEP-01
PUB DATE 1993-06-00
NOTE 49p.; "In collaboration with Eugene J. Johnson, Charles Lewis, Nancy A. Mead; Data Anslyses by Duanli Yan."
CONTRACT SED-9255369; IAD-91-0222
AVAILABLE FROM Educational Testing Service, Center for the Assessment of Educational Progress, Rosedale Road, Princeton, NJ 08541-0001.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Standards; Elementary Secondary Education; Foreign Countries; Interviews; Mathematics Curriculum; *Mathematics Instruction; *National Competency Tests; Research Methodology; *Student Evaluation; Test Interpretation
IDENTIFIERS Educational Testing Service; *International Assessment of Educational Progress; *National Assessment of Educational Progress

ABSTRACT

This study investigates a linking of the 1991 International Assessment of Educational Progress (IAEP) and the 1992 National Assessment of Educational Progress (NAEP) mathematics assessments. Data to allow such a linking were collected in 1992 from students in the United States who were administered both instruments. Modeling was done using a regression analysis and that model was used as the basis for projecting IAEP scores from non-United States countries onto the NAEP scale. This study focuses on the percentage of students from the IAEP countries who were predicted to fall above the three National Assessment Governing Board (NAGB) achievement levels and investigates four sources of error. These sources of error pertain to not having or not knowing: (1) the true relationship between the IAEP and NAEP assessments; (2) results for the entire IAEP population; (3) simple random samples of students; and (4) the true proficiency level of every student. Results of the study were very encouraging. The relationship between the IAEP and NAEP assessments was quite strong and could be modeled well. (Contains 16 references.) (Author/DDR)

Toward World-Class Standards

A Research Study Linking International and National Assessments

Peter J. Pashley ▪ Gary W. Phillips

in collaboration with

Eugene J. Johnson ▪ Charles Lewis ▪ Nancy A. Mead

Data Analyses by Duanli Yan

June 1993

Educational Testing Service

Educational Testing Service, (ETS) is a private, not for profit corporation devoted to measurement and research, primarily in the field of education. It was founded in 1947 by the American Council on Education, the Carnegie Foundation for the Advancement of Teaching, and the College Entrance Examination Board.

The Center for the Assessment of Educational Progress (CAEP) is a division of ETS devoted to innovative approaches to the measurement and evaluation of educational progress. The present core activity of CAEP is the administration of the National Assessment of Educational Progress (NAEP), under contract from the U.S. Department of Education. CAEP also carries out related activities, including the International Assessment of Educational Progress (IAEP), state assessments, and special studies such as the National Science Foundation-supported Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics.

The work upon which this publication is based was performed under a subcontract agreement with Westat, Inc. pursuant to prime contract No. SED-9255369 of the National Science Foundation, and additional funding was provided by the U.S. Department of Education through Interagency Agreement No. IAD-91-0222. Supplementary funds were provided by the Carnegie Corporation of New York.

This report, No. 24-CAEP-01, can be ordered from the Center for the Assessment of Educational Progress at Educational Testing Service, Rosedale Road, Princeton, New Jersey 08541-0001.

Educational Testing Service is an equal opportunity/affirmative action employer.

CONTENTS

Executive Summary 1

Introduction 3

Methodology 11

Validation 21

Results 24

Conclusions 32

References 35

Appendix I: Data Collection 37

Appendix II: Notation 43

Appendix III: Computer Packages 45

Appendix IV: TBLT Results 47

EXECUTIVE SUMMARY

The allure of linking large- and small-scale assessments has been growing steadily among educators, researchers, and policymakers. While much has been written on the theoretical implications of such endeavors, little empirical evidence is currently available to guide those interested in conducting these linkings. This study attempts to rectify this situation.

The study investigated a linking of the 1991 International Assessment of Educational Progress (IAEP) and the 1992 National Assessment of Educational Progress (NAEP) mathematics assessments. Sample data to allow such a linking were collected in 1992 from U.S. students who were administered both instruments. The relationship between mathematics proficiencies yielded by these two assessments was then modeled by way of a regression analysis. This model was used in turn as a basis for projecting IAEP scores from non-U.S. countries onto the NAEP scale. The study focused on the percentage of students from the IAEP countries predicted to fall above the three National Assessment Governing Board (NAGB) achievement levels.

While estimating percentages above a NAGB achievement level is relatively straightforward, the real challenge was to assess all the related components of error that are possibly associated with such estimates. This study investigated four sources of error. These result from not having or not knowing the following: 1) the true relationship between the IAEP and NAEP assessments, 2) results for the entire IAEP populations, 3) simple random samples of students, and 4) the true proficiency level of every student. These components were quantified to derive standard errors corresponding to the estimates of percentages, and then used to construct confidence intervals related to the estimates for each IAEP country and NAGB achievement level.

The results of this study were very encouraging. The relationship between the IAEP and NAEP assessments was quite strong and could be modeled well. The largest components of error found were related to the uncertainty of estimating population values based on a (non-simple random) sample. The derived confidence intervals exhibited an average range of about 3 percent. Results from a cross-validation study indicated that the proposed methodology is quite stable, even with relatively small sample sizes.

This study should provide good news for policymakers, among others, who are interested in linking large- and small-scale assessments. On the other hand, the results presented here should be put into context. To begin with, the IAEP and NAEP mathematics assessments are fairly similar in their construction and scoring. Also a number of assumptions (or leaps of faith) must be made before the results can be taken seriously. The first is that the relationship between the IAEP and NAEP assessments observed in the U.S. linking sample is also applicable to other countries. The second is that this same relationship based on 1992 data also holds for the 1991 assessment. The third assumption that must be made is that other unexplored sources of error, such as motivation levels, would not significantly change the results.

So while the path to linking assessments is now better understood, investigators should still proceed with caution.

INTRODUCTION

There is a lot of interest among educators, researchers, policymakers, and the general public in how the American educational system compares to those in other countries. One indicator that is often looked at in these comparisons is outcome measures of what students have learned in school. Recent studies, such as the 1991 Reading Literacy Study conducted by the International Association for the Evaluation of Educational Achievement (IEA) and the 1991 International Assessment of Educational Progress (IAEP) in mathematics and science conducted by Educational Testing Service, have heightened the interest in such international comparisons. In the 1991 IAEP study, one state, Colorado, drew a large enough sample to compare itself to all 20 participating countries.

In 1989, the United States made it clear that it was very serious about international comparisons, especially in the area of mathematics and science. The nation's 50 Governors, along with the President, held a National Education Summit and adopted six education goals. The fourth goal states that by the year 2000, "U.S. students will be first in the world in science and mathematics achievement" (National Education Goals Panel, 1991, p. 16). Since that time a variety of approaches have been suggested on how data might be collected that could help the nation monitor progress toward that goal. One of the most ambitious efforts will be the 1995 IEA Third International Mathematics and Science Study (TIMSS); almost 60 countries have expressed an interest in participating in the study.

Although the TIMSS study may be the best source of data to monitor progress toward the fourth educational goal during this decade, the results from the first phase of the study will not be available until 1996 or 1997.

In addition to expressing an interest in international comparisons, policymakers have recently been recommending linking assessments. The idea behind this proposal is that combining data collected across testing programs might result in considerable cross-fertilization of information. For example, in their January 1992 report the

National Council on Education Standards and Testing (NCEST) recommended that groups or clusters of states should "adopt assessments linked to national standards" so that different assessments can "produce comparable results in the attainment of the standards" (1992, p 30).

The National Assessment Governing Board (NAGB) is another example of a policy group recommending such linking strategies. In its 1993 paper, which discusses the future of the National Assessment of Educational Progress (NAEP), the board recommends that "States, school districts, and schools should be permitted to use NAEP to link the results of their local assessments with national and international results...using NAEP in this way poses significant technical challenges; thus, research and development in this area should continue" (1993, p 8). NAGB is in the process of drafting a policy that will provide procedural and technical guidelines that should be followed in linking other assessments to NAEP. Also, the Education Information Advisory Committee (EIAC) of the Council of Chief State School Officers (CCSSO) has endorsed the use of NAEP as an anchor to which local tests might be linked, but they have also warned that "technical difficulties are numerous and not easily overcome" (National Assessment Governing Board, 1993, p. 8).

NAEP is a Congressionally mandated assessment of public and private school students. It is currently conducted biennially in grades 4, 8, and 12 and at ages 9, 13, and 17. Representative samples of students are administered achievement tests in a variety of subject areas. For example, in 1990 students were tested in reading, mathematics, and science. In 1992 they were tested in reading, mathematics, and writing. Background questionnaires are also administered to students, teachers, and principals in an effort to provide contextual information for student learning. Starting in 1990 Congress also authorized a voluntary Trial State Assessment. In 1990 37 states, territories, and the District of Columbia participated in the Trial State Assessment. The trial was expanded to 44 education jurisdictions in 1992.

The impetus for this IAEP-NAEP link study came from the efforts of the National Center for Education Statistics (NCES) to begin the research and development work in this area and to pilot the linking of NAEP to other assessments. In 1992-93, there were two opportunities to conduct studies on how NAEP might be linked to other testing programs. One occurred when the General Assembly of the Commonwealth of Kentucky passed the Kentucky Education Reform Act. The legislation called for a "NAEP-like test" that would be "an interim testing program to access student skills in reading, mathematics, writing, science, and social studies in

grades four, eight, and twelve." Kentucky used the NAEP framework to guide development for each of these subject areas resulting in tests with content similar to NAEP. Results of a study that links the Kentucky tests to NAEP in reading, mathematics, and writing will be available in 1993.

The second opportunity to link NAEP to other assessments occurred with the 1991 IAEP. The IAEP was conducted by Educational Testing Service with funds from the National Center for Education Statistics and the National Science Foundation. Representative samples of 9- and 13-year-old students were tested in mathematics and science in 20 countries. Those countries decided to adopt the 1990 NAEP objectives in mathematics as a blueprint for the construction of the IAEP mathematics assessment. Therefore, there was substantial content overlap between NAEP and the IAEP. Even though there were differences in the target population and timing between the IAEP and NAEP (the IAEP assessed age samples in 1991 whereas the NAEP assessed grade samples in 1992), it was felt there was enough overlap to do an experimental linking study. This report provides the results of that study.

More specifically, by linking the IAEP scale to the NAEP scale it is possible to predict the percentages of 13-year-olds in each of the 20 countries that participated in the 1991 IAEP in mathematics who would have performed at or above each of the three achievement levels established by the NAGB for U.S. students. These predications can then be compared with actual performance of U.S. eighth graders in public schools in the 1990 and 1992 NAEP mathematics assessments with respect to these same criteria.

The NAGB achievement levels are the outcome of a standard setting process that established three points at each of three grade levels along the NAEP mathematics scale (which ranges from 0 to 500) that represent basic, proficient, and advanced levels of performance at each grade level. These levels are defined in FIGURE 1.

FIGURE 1: Description of Mathematics Achievement Levels for Basic, Advanced, and Proficient Eighth Graders

The five NAEP content areas are: (1) numbers and operations, (2) measurement, (3) geometry, (4) data analysis, statistics, and probability, and (5) algebra and functions. Skills are cumulative across levels -- from Basic to Proficient to Advanced.

Basic 256	Eighth-grade students performing at the basic level should exhibit evidence of conceptual and procedural understanding in the five NAEP content areas. This level of performance signifies an understanding of arithmetic operations -- including estimation -- on whole numbers, decimals, fractions, and percents.

Eighth graders performing at the basic level should complete problems correctly with the help of structural prompts such as diagrams, charts, and graphs. They should be able to solve problems in all NAEP content areas through the appropriate selection and use of strategies and technological tools -- including calculators, computers, and geometric shapes. Students at this level also should be able to use fundamental algebraic and informal geometric concepts in problem solving.

As they approach the proficient level, students at the basic level should be able to determine which of available data are necessary and sufficient for correct solutions and use them in problem solving. However, these eighth graders show limited skill in communicating mathematically.

Proficient 294	Eighth-grade students performing at the proficient level should apply mathematical concepts and procedures consistently to complex problems in the five NAEP content areas.

Eighth graders performing at the proficient level should be able to conjecture, defend their ideas, and give supporting examples. They should understand the connections between fractions, percents, decimals, and other mathematical topics such as algebra and functions. Students at this level are expected to have a thorough understanding of basic level arithmetic operations -- an understanding sufficient for problem solving in practical situations.

Quantity and spatial relationships in problem solving and reasoning should be familiar to them, and they should be able to convey underlying reasoning skills beyond the level of arithmetic. They should be able to compare and contrast mathematical ideas and generate their own examples. These students should make inferences from data and graphs; apply properties of informal geometry; and accurately use the tools of technology. Students at this level should understand the process of gathering and organizing data and be able to calculate, evaluate, and communicate results within the domain of statistics and probability.

Advanced 331	Eighth-grade students performing at the advanced level should be able to reach beyond the recognition, identification,a nd application of mathematical rules in order to generalize and synthesize concepts and principles in the five NAEP content areas.

Eighth graders performing at the advanced level should be able to probe examples and counterexamples in order to shape generalizations from which they can develop models. Eighth graders performing at the advanced level should use number sense and geometric awareness to consider the reasonableness of an answer. They are expected t u se abstract thinking to create unique problem-solving techniques and explain the reasoning processes underlying their conclusions.

Recent papers by Linn (in press) and Mislevy (1992) have contributed substantially to the literature on linking assessments. Both authors outline linking strategies, discuss their strengths and weaknesses, and clarify the terminology. The IAEP-NAEP study took into account the concepts in these reports. Therefore, it will be useful to review the range of linking procedures so that the benefits and limitations of the one used in this study can be better understood.

The reports by Linn and Mislevy outline four methods of linking assessments: equating, calibration, projection, and moderation. The four methods are ordered according to the degree to which various assumptions must be met.

Equating procedures are employed when we have two alternative forms of the same test. Examples include new forms of a driver's license test, or new forms of the SAT, ACT, or the Advanced Placement Tests. After one test is equated to the other, the forms are interchangeable, and it doesn't matter to the examinee which form of the test he/she is taking. The examinee would have received the same test score (or the average) regardless of which test was administered. Two tests can be equated when they measure the same thing and are equally reliable. When these conditions are met statistical linking yields its maximum benefits. These include:

- a single correspondence table provides conversions between both tests
- conversions for group distributions also apply to individuals
- the need for checks on stability over subgroups, context, and time decreases

Calibration procedures are used when two tests measure the same thing but one test is longer or more reliable than the other. Under these conditions the two tests cannot be equated, but they can be calibrated to a common scale. NAEP uses calibration procedures across test booklets within an assessment and across time in order to maintain a common 0-500 scale. Many norm-referenced test publishers use calibration procedures to create a common scale across grades. The paper by Linn (1993) provides a good example of calibration that illustrates why less reliable tests cannot be equated to more reliable tests.. Imagine a basketball coach who wants to select players who shoot with at least 75 percent accuracy. The coach uses a short form of a test consisting of four attempts with player 1 and a long form of a test consisting of 20 attempts with player 2. If both players have a true accuracy of 50 percent, player 1 will have .31 chance of reaching a 75 percent level of accuracy in four shots whereas player 2 will have less than a .01 chance of reaching this level in 20 shots. This occurs because four attempts provide less reliable data than 20 attempts. It would be a mistake to compare the results of these two tests without taking into account these differences in reliability.

Projection procedures (usually regression analysis) are used when both of the assumptions of equating are relaxed. With projection procedures the two tests need not measure the same trait nor be equally reliable. The goal of projection is to *predict* the scores on one test from the scores on the other test. Both calibration and projection result in a linking of scores between two tests. In the case of calibration, the two tests are linked via a common scale, and in the case of projection, they are linked through a prediction equation. However, in both cases, the statistical benefits (mentioned above) of equating are lost. This is the statistical price we pay for not meeting the two assumptions required for equating.

For both calibration and projection procedures, we have to do a lot more work to make sure the inferences from our analysis are valid. For example, there is no longer a single conversion table between the two tests. The translation from test x to test y may be different from the translation from y to x. Also, the conversion procedures needed for individuals may be different than the ones needed for groups. The contribution of equating error to the equated scores may be unequal throughout the score range. Generally, the equated scores will have more error in the tails of the distribution. Finally, the conversion procedures will have to be checked to make sure they apply to subgroups, and that they hold up across different contexts and time.

Moderation is a procedure used commonly in European examination systems but less often in the United States. In moderation, the scores of one test are adjusted (usually using the formulas of equating), but there is no claim that the data meet the two assumptions required for equating. When statistical formulas are used the procedure is called **statistical moderation.** When direct judgments are made about the comparability of performance levels on different assessments the procedure is called **social moderation.** For example, "social moderation might involve the independent rating of a teacher's classroom by other teachers within the same school or by teachers and expert raters from other schools...Differences in ratings would then be discussed in an effort to achieve consensus" (Linn, in press).

Mislevy has argued that moderation techniques are really outside the realm of statistical inference: "Moderation should not be viewed as an application of principles of statistical inference, but as a way to specify the rules of the game. It can yield an agreed upon way of comparing students who differ qualitatively, but it doesn't make information from tests that aren't built to measure the same thing as if they did. An arbitrarily determined operational definition of comparability must be defended" (1992, p. 72).

From the four approaches outlined by Mislevy, this study uses the projection procedures. In doing so, we acknowledge that NAEP and IAEP have not been equated, but instead the NAEP scores have been predicted from the IAEP scale. Therefore, the statistical benefits that result from equating have not been obtained in this study. In reviewing this research study the reader should keep the following limitations in mind.

- The equation used to predict NAEP results from the IAEP study is different from the equation needed to predict IAEP results from NAEP. Because this study was interested in how students in other countries would do on the NAEP test based on predictions from the IAEP, only the first set of prediction equations is used.

- The prediction equation developed in this study was used to estimate performance of groups of students rather than individuals.

- The standard errors provided in this study are more complicated than those obtained from most surveys. The standard errors are affected by a) linking (or regression estimation) error, b) sampling error, c) design effects, and d) measurement error. The standard errors indicted in the tables are the sum of these four components.

- International predictions are provided for 13-year-olds in public and private schools who participated in the 1991 IAEP mathematics assessment. Actual U. S. performance is provided for eighth-grade students in public schools who participated in the NAEP mathematics assessment in 1990 and 1992.

We hope that the results of this study will be used to illustrate the linking technology that others might use in linking assessments to NAEP. We also hope that others will benefit from this initial trial effort and make improvements in the future.

METHODOLOGY

To achieve the goals of this study, we needed a methodology for estimating the proportion of an IAEP population falling above a cut-score on the NAEP scale, along with the associated standard error. The resulting technology could then be applied to different IAEP populations and cut-scores (such as NAGB achievement levels). Many approaches to this problem can be taken. One method, based on a regression of NAEP scores on IAEP values, will be examined in detail in this section.

The preferred estimation approach, from which the results found in the next section were derived, can be summarized in the following steps: obtain linking sample results, compare the 1991 and 1992 IAEP scales, derive IAEP and NAEP plausible proficiency values for the linking sample, model the relationship between the IAEP and NAEP linking sample results, produce estimates of cut-score proportions, and calculate associated standard errors to create confidence intervals. These steps are described below. For each, a brief non-technical summary is given (in italics) followed by some technical details.

More details are given in the appendices: a complete description of the linking sample and how it was chosen is found in Appendix I; the notation that will be used in the technical areas is outlined in Appendix II; the computer packages used to produce the results are covered in Appendix III; and some scaling results are given in Appendix IV. Further information on NAEP and IAEP scaling can be found in their associated technical reports (i.e., Johnson & Allen, 1992; J.-G. Blais, 1992, respectively).

Obtain Linking Sample Results

In order to establish a link between the IAEP and NAEP assessments, a sample of 1,609 U.S. grade eight students were assessed with both instruments in 1992.

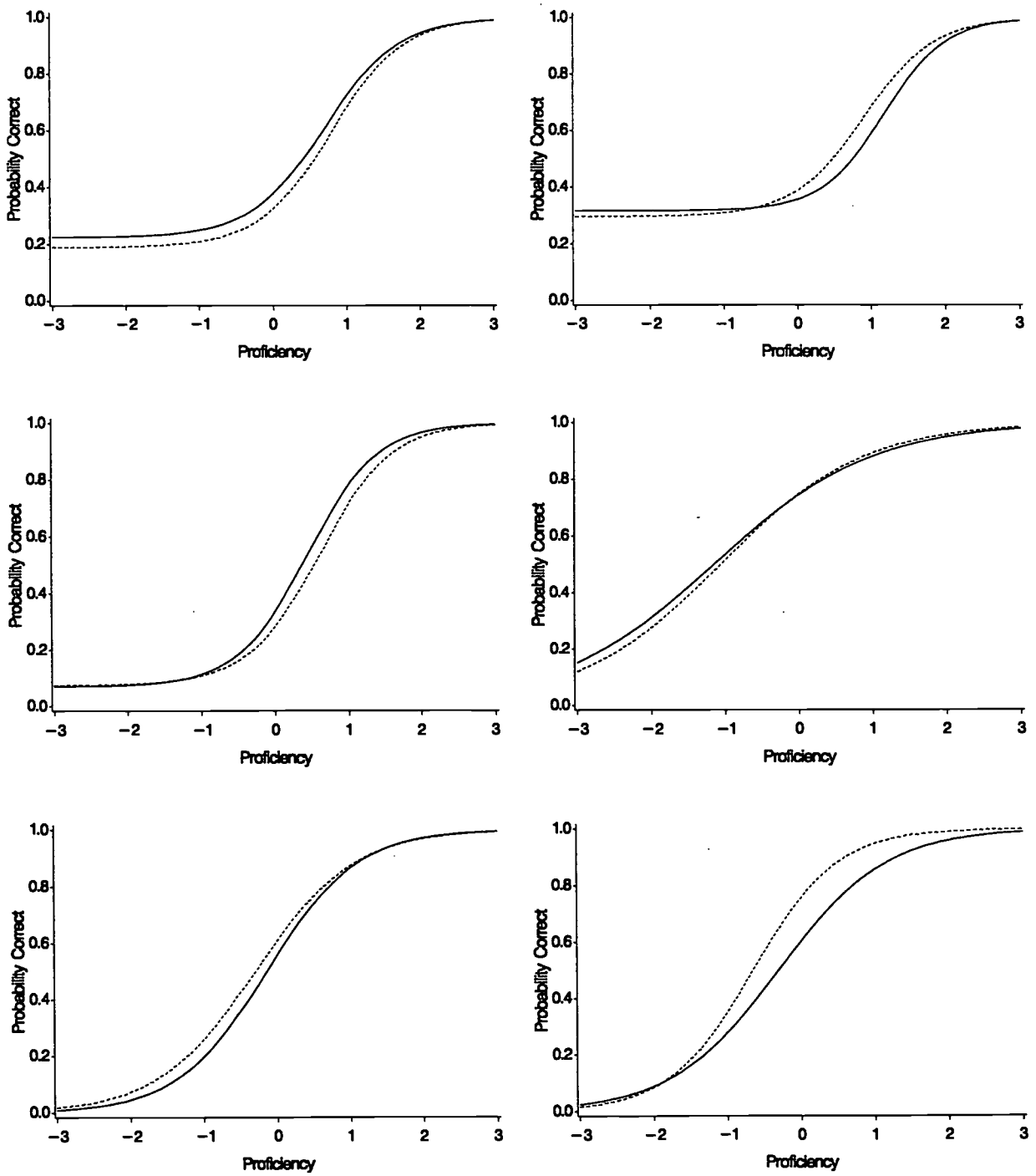
that students (eighth-graders/13-year-olds) who were assessed in a NAEP mathematics session could be re-assessed with the IAEP instrument. The representativeness of this school sample closely paralleled that of the original national sample. Note that sampling weights were not used in the calculation of the results cited in this study. The decision not to include sampling weights was made because they were deemed not essential in establishing a valid linking equation and because of the concern that this added complexity might obscure the impact of the new methodology presented in this study.

Compare 1991 and 1992 IAEP Scales

While the IAEP/NAEP linking sample was assessed in 1992, the last full IAEP study took place in 1991. In addition, while the 1991 items were calibrated from a "super-sample" selected from all the participating countries, the 1992 linking sample consisted of only U.S. students. Therefore, 1992 IAEP item characteristics (i.e., parameters) were estimated from the 1992 linking sample and compared to those obtained from the 1991 study. Results suggested that there were no significant discrepancies between items independently calibrated under these two conditions.

The 1992 IAEP linking sample was used to re-calibrate the IAEP items using BILOG. Procedures similar to those used in the initial 1991 IAEP calibration were followed (see J.-G. Blais, 1992). A sample of the item response curves estimated from these two calibrations are shown in FIGURE 2. To try and improve the similarity between these curves, and related proficiency scales, the program TBLT, which adjusts proficiency scales to approximately line up the test characteristic curves, was utilized. Results from this analysis are given in Appendix IV. Due to the small discrepancies between the item response curves and the trivial effect of the TBLT program, we decided to use the original 1991 item response parameter estimates in the calculation of 1992 IAEP proficiency estimates.

FIGURE 2: ***A Sample of IAEP Mathematics Item Response Curves for 1991 IAEP Super-population Sample (solid line) and 1992 IAEP Linking Sample (dashed line)***



Derive Proficiencies for the Linking Sample

Two independent sets of five imputed proficiency values, corresponding to the IAEP (mathematics) and NAEP (composite mathematics) instruments, were derived for every student in the linking sample.

In NAEP, each assessed student typically takes too few items to permit a reliable estimate of individual proficiencies. Instead, NAEP uses a procedure which generates a predictive distribution of potential scale scores for each individual. This predictive distribution is based on the student's responses to the cognitive items and on the student's status on several hundred "conditioning variables" based on background characteristics. Drawn from this distribution are imputed proficiencies (called plausible values) that are used in the place of individual proficiencies for analysis. The plausible values provide appropriate estimates of subpopulation proficiency distributions and account for the error due to imprecision of individual measurement (see Mislevy, Johnson, and Muraki, 1992). The same technology was used for the IAEP data, although the error due to imprecision of individual measurement was less important because each student responded to more items than in the NAEP.

Using previous 1991 IAEP and national 1992 NAEP item parameters, two independent sets of five imputed plausible values were derived using the computer program MGROUP. Similar background information was used as conditioning variables. The NAEP plausible values were obtained for each of five content area scales within mathematics, which were then averaged into an overall mathematics composite. The IAEP proficiency plausible values were converted into deviation scores (see Appendix II).

Regress NAEP Scores on IAEP Values

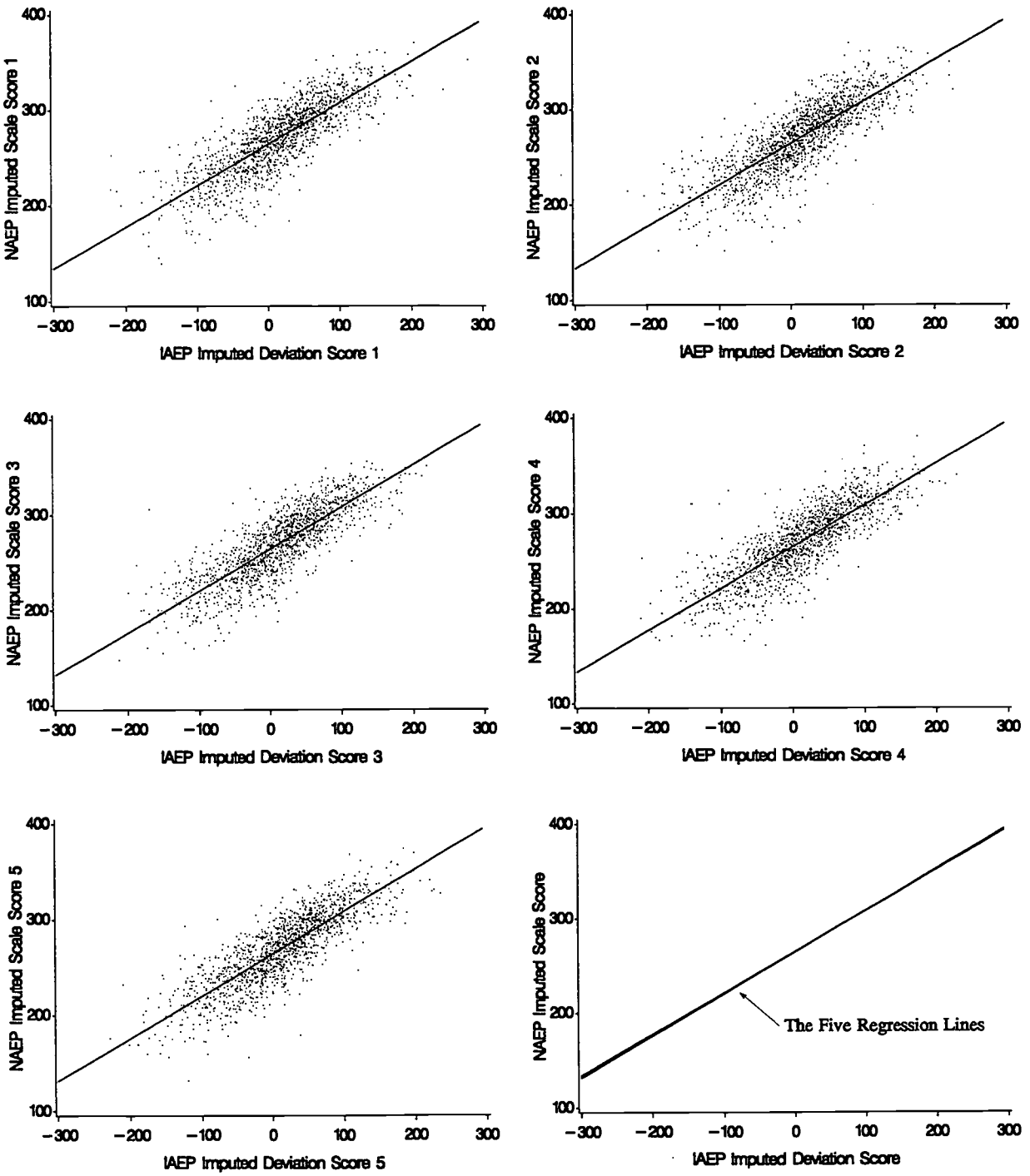
Based on the linking sample results, the relationships between the IAEP and NAEP imputed values were investigated. Simple linear regressions of NAEP on IAEP were found to adequately model these relationships.

IAEP deviation scores, for each student in the linking sample. From these five pairings, five linear regression equations were derived. The regression results are shown in TABLE 1 below. FIGURE 3 illustrates the fit of the regression lines to the data and the small differences among the five lines.

TABLE 1: ***Parameter Estimates and Root Mean Square Errors From the Regressions of 1992 NAEP Plausible Values on 1992 IAEP Plausible Values***

Imputation Pairing	Parameter Estimates		Root Mean Square Error (RMSE)
	Slope (β)	Intercept (α)	
1	270.0	.44	20.7
2	269.9	.44	21.2
3	269.6	.44	20.5
4	270.3	.44	20.2
5	269.8	.44	20.1

FIGURE 3: ***Regressions of 1992 NAEP Plausible Values on 1992 IAEP Plausible Values***



Calculate Estimates of Proportions

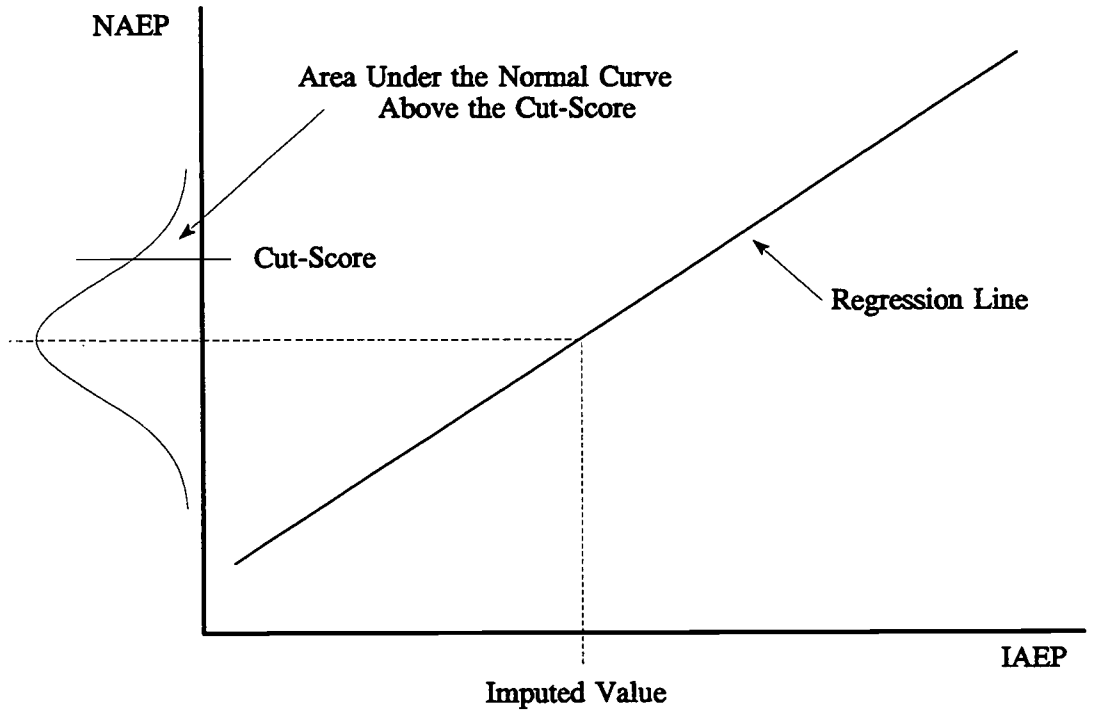
First consider the simplified problem of calculating the probability of falling above a NAEP cut-score (or NAGB achievement level) for one 1991 IAEP examinee, based on one imputed value. Using one 1992 IAEP/NAEP calibrated regression line, one NAEP score can be estimated. Assuming this score is normally distributed, the probability that (with repeated sampling) such scores will fall above a NAEP cut-score can be calculated. This situation is illustrated in FIGURE 3. Then in a similar fashion, four other probabilities can be derived for this same student, based on the four other imputed values and linear regressions. Five such probabilities can then be derived for each student in a IAEP sample. The average of all of these probability estimates provides an estimate of the proportion of an IAEP population that will fall above a NAEP cut-score.

The probability that an examinee in an application sample will fall above a particular cut-score on the NAEP scale can be estimated by

$$\hat{p}_{aj(k)} = 1-\Phi\left[z_{aj(k)}\right]$$

the area under the normal curve above the cut-score in FIGURE 4. This estimate is based on one imputed value and given the usual normality and homogeneity of variance assumptions. (Again, for details on the notation, see Appendix II.) Then a point estimate of the proportion of an IAEP population falling above a cut-score c on the NAEP scale is $\hat{p}_{a\cdot(\cdot)}$ (i.e., the average proportion across the sampled examinees and imputations).

FIGURE 4: The Probability of Falling Above a NAEP Cut-score Given an Imputed IAEP Value



Derive Standard Errors

Point estimates can be affected by many types of error. In this study, four of these sources were accounted for in the standard error calculations. In particular, we considered errors associated with regression estimation, sample-to-population estimation, design effects, and measurement.

We will begin by simplifying the problem by assuming that there exists no measurement error. (This component will be addressed later on.) If the IAEP and NAEP tests were perfectly reliable, then the five imputed values would be equal. Under this assumption, we need only investigate the error associated with the average proportion across one set of imputed values (i.e., $\hat{p}_{a\cdot(k)}$).

Regression estimation errors. For the moment, also assume that we are only interested in the estimated proportion for a simple random sample from an application population. In this simplified problem, we must still take into account the error associated with estimating the regression of NAEP on IAEP. We will denote this variance by $\sigma_r^2(\hat{p}_{a\cdot(k)} \mid a)$, where

$$\hat{p}_{a\cdot(k)} = 1 - \frac{\sum_{j=1}^{n_a} \Phi[z_{aj(k)}]}{n_a}$$

and n_a is the application sample size.

As the usual linear variance formulae are not applicable to this situation, a Taylor series approximation was used:

$$\hat{\sigma}_r^2(\hat{p}_{a\cdot(k)} \mid a) = \frac{1}{n_e}\left[\text{Ave}(\phi[z_{aj(k)}])\right]^2 + \frac{1}{\sum_j x_{ej(k)}^2}\left[\text{Ave}(x_{aj(k)}\phi[z_{aj(k)}])\right]^2 + \frac{1}{2(n_e-2)}\left[\text{Ave}(z_{aj(k)}\phi[z_{aj(k)}])\right]^2 \quad (1)$$

The three components in (1) correspond to the sampling variance of the estimated intercept, slope, and residual variance.

Sample-to-population estimation errors. Now assume that the regression line is given and we are interested in the error associated with estimating a population proportion from a simple random sample. An unbiased estimate of this quantity (again using a Taylor series approximation) is

$$\hat{\sigma}_s^2(\hat{p}_{a\cdot(k)} \mid \hat{\alpha}_{(k)}, \hat{\beta}_{(k)}, RMSE_{(k)}) = \frac{1}{n_a}\text{Var}(\hat{p}_{aj(k)}) - \frac{1}{n_e n_a}\text{Var}(\phi[z_{aj(k)}]) - \frac{1}{n_a\sum_{j=1}^{n_e} x_{ej(k)}^2}\text{Var}(x_{aj(k)}\phi[z_{aj(k)}]) - \frac{1}{2(n_e-2)n_a}\text{Var}(z_{aj(k)}\phi[z_{aj(k)}]) \quad (2)$$

Design effects. Assuming statistical independence, the unconditional variance associated with $\hat{p}_{a\cdot(k)}$ can be estimated by

$$\hat{\sigma}^2\left(\hat{p}_{a\cdot(k)}\right) = \hat{\sigma}_r^2\left(\hat{p}_{a\cdot(k)} \mid a\right) + \hat{\sigma}_s^2\left(\hat{p}_{a\cdot(k)} \mid \hat{\alpha}_{(k)},\ \hat{\beta}_{(k)},\ RMSE_{(k)}\right) \tag{3}$$

This variance estimate is appropriate when a simple random sample of examinees is selected from an IAEP population. If this is not the case, then the Var(·) values may be too small. Analysis of the data from the IAEP assessment indicates that the variances assuming simple random sampling are too small by a factor of about 2. Adjusting the variances by this factor (called a design effect) leads to a final variance estimate, assuming no measurement error of

$$\begin{aligned}\hat{\sigma}^2\left(\hat{p}_{a\cdot(k)}\right) = {} & \frac{1}{n_e}\left\{\left[\mathrm{Ave}(\phi[z_{aj(k)}])\right]^2 - \frac{2}{n_a}\mathrm{Var}(\phi[z_{aj(k)}])\right\} + \\ & \frac{1}{\sum_j x_{ej(k)}^2}\left\{\left[\mathrm{Ave}(x_{aj(k)}\phi[z_{aj(k)}])\right]^2 - \frac{2}{n_a}\mathrm{Var}(x_{aj(k)}\phi[z_{aj(k)}])\right\} + \\ & \frac{1}{2(n_e-2)}\left\{\left[\mathrm{Ave}(z_{aj(k)}\phi[z_{aj(k)}])\right]^2 - \frac{2}{n_a}\mathrm{Var}(z_{aj(k)}\phi[z_{aj(k)}])\right\} + \\ & \frac{2}{n_a}\mathrm{Var}\left(\hat{p}_{aj(k)}\right)\end{aligned} \tag{4}$$

Measurement error. Finally, measurement error must be accounted for. An approach inspired by the typical one-way analysis-of-variance design was applied in this case. That is, variance from between and within imputations were added together to estimate the variance associated with the overall point estimate:

$$\hat{\sigma}^2\left(\hat{p}_{a\cdot(\cdot)}\right) = \mathrm{Ave}\left[\hat{\sigma}^2\left(\hat{p}_{a\cdot(k)}\right)\right] + \mathrm{Var}\left(\hat{p}_{a\cdot(k)}\right) \tag{5}$$

Then an approximately 95 percent confidence interval for the proportion of the specified country population falling above a cut-score c on the NAEP scale is

$$\hat{p}_{a\cdot(\cdot)} \pm 2\hat{\sigma}\left(\hat{p}_{a\cdot(\cdot)}\right)$$

VALIDATION

Cross-Validation

Projection procedures are not guaranteed to work equally well for all subpopulations for which projections might be desired. For example, regression procedures, such as used for this study, tend to capitalize on features of the sample at hand, producing predictions that will generally be better for the sample used to construct the equation (the U.S. linking sample) than they will be for other samples (such as the other IAEP populations).

As one way to empirically assess the accuracy of the proposed methodology, we conducted a cross-validation study, based on the 1992 IAEP/NAEP linking sample. This evaluation proceeded as follows. The entire linking sample was first split into half-samples. Then two sets of NAEP on IAEP regression lines were independently calibrated (for each half-sample). Estimated confidence intervals were then derived by using one half-sample as a "linking sample" and the other as an "application sample," and then proceeding as outlined in the previous section. The cut scores used were the three NAGB achievement levels: Basic (256), Proficient (294), and Advanced (331). Then a second set of confidence intervals could be calculated by switching the half-sample designations.

One set of results is shown in TABLE 2. One of the half-samples in this case was constructed by randomly selecting one third of the examinees scoring above the IAEP median and two-thirds of those scoring below the median. Those students not chosen were then placed in the other half-sample. This allocation of examinees yielded a significant difference in the distribution of proficiencies across the two half-samples. This difference is reflected in the results listed in TABLE 2. Proportions have been transformed to percentages to maintain comparability with NAEP reporting.

TABLE 2: Cross-Validation Results

Half-Sample	NAGB Achievement Level	Estimated Confidence Interval	Actual Percentage At or Above NAGB Achievement Level
1	Basic (256)	55.1 - 60.8	56
	Proficient (294)	16.3 - 20.6	18
	Advanced (331)	1.9 - 4.0	2
2	Basic (256)	70.4 - 75.4	75
	Proficient (294)	28.0 - 33.7	36
	Advanced (331)	3.9 - 6.9	5

Note that the percentages labeled "actual" in TABLE 2 are based on the "application sample" and, therefore, are only estimates of the true population percentages. (They represent the average over the five imputations.) As such, they also contain some error that should be taken into account when considering the outcomes. Even with this additional source of error, the results are quite good. Only one "actual" percentage falls outside the corresponding estimated confidence interval (i.e., for half-sample #2, NAGB achievement level 294). Results from randomly divided half-samples yielded similar results. This adds credibility to using this procedure to project results for widely variable proficiency distributions, such as seen for the IAEP populations.

Estimating 1992 U.S. Percentages

As another way to empirically assess this methodology, we used the entire linking sample to estimate confidence intervals for the 1992 U.S. proportions. These were then compared to "actual" percentages that had been previously calculated from the full 1992 NAEP study (see Mullis *et al*, 1993, Table 1.1, p. 64). As shown in TABLE 3, all three "actual" percentages fall within or close to the estimated confidence intervals. Given that there is sampling error associated with these "actual" percentages (for example, the standard error for the percentage at or above the basic level is 1.1), this result provides further evidence that the procedure is providing good predictions.

TABLE 3: Linking Sample Results

NAGB Achievement Level	Estimated Confidence Interval	Actual Percentage At or Above NAGB Achievement Level
Basic (256)	63.3 - 68.0	63
Proficient (294)	23.2 - 27.4	25
Advanced (331)	3.5 - 5.3	4

RESULTS

As a result of the statistical work described in the prior sections, it was possible to estimate the percentages of students for the 15 IAEP "comprehensive" populations and the six IAEP populations with "exclusions or low participation" that would perform at or above the three NAGB achievement levels. To emphasize the uncertainty associated with these percentage estimates, confidence (or percentage) intervals are given in TABLES 4 and 5. These intervals were formed by adding and subtracting two times the corresponding standard error (as outlined in the methodology section). The results are approximate 95 percent confidence intervals. These intervals are illustrated graphically in FIGURES 5 and 6. These results can be compared with those for the United States, broken down by regions, states, and territories, found in TABLES 6 and 7. These U.S. results are based on NAEP assessments of students in grade 8 in public schools conducted in 1990 and 1992. Note that the IAEP assessment was conducted in 1991 with 13-year-old students in public and private schools, so direct comparisons are not advised.

Substantially, the results indicate that the Asian countries participating in the IAEP would be predicted to do very well if they had taken the NAEP test. As an example, Korea and Taiwan were predicted to have 5 to 7 percent and 9 to 12 percent, respectively, of their 13-year-old students scoring at or above the advanced achievement level. This compares with 2 percent of U.S. public school eighth-grade students at this level in 1990, and 3 percent in 1992.

TABLE 4: ***Percentage (Confidence) Intervals At or Above NAGB Achievement Levels for 1991 IAEP Mathematics for Comprehensive Populations -- Age 13 Public and Private Schools***

Comprehensive Populations	Percentage (Confidence) Intervals of Students At or Above NAGB Achievement Levels		
	Basic (256)	Proficient (294)	Advanced (331)
Korea	78.7 - 82.5	34.6 - 39.3	05.3 - 07.5
Taiwan	75.9 - 80.1	38.2 - 43.1	08.8 - 12.0
Switzerland 15 Cantons	82.7 - 85.1	31.6 - 34.6	02.8 - 03.9
Soviet Union Russian-speaking Schools in 14 Republics	77.9 - 81.4	28.1 - 31.9	02.4 - 03.7
Hungary	74.6 - 78.7	26.6 - 30.6	02.6 - 04.0
France	70.7 - 74.6	21.4 - 25.0	01.5 - 02.5
Emilia-Romagna, Italy	70.2 - 74.2	20.2 - 23.9	01.2 - 02.1
Israel Hebrew-speaking Schools	70.5 - 74.5	19.8 - 23.3	01.1 - 01.9
Canada	68.9 - 71.6	18.2 - 20.6	01.1 - 01.7
Scotland	66.9 - 71.0	17.8 - 21.2	01.0 - 01.8
Ireland	65.5 - 69.8	17.5 - 20.9	01.0 - 01.9
Slovenia	62.3 - 66.7	14.4 - 17.5	00.6 - 01.3
Spain Spanish-speaking Schools except in Cataluña	57.5 - 61.8	10.4 - 13.0	00.3 - 00.7
United States	55.6 - 60.5	12.6 - 15.9	00.8 - 01.7
Jordan	35.4 - 40.2	04.6 - 06.5	00.1 - 00.3

TABLE 5: ***Percentage (Confidence) Intervals At or Above NAGB Achievement Levels for 1991 IAEP Mathematics for Populations with Exclusions or Low Participation – Age 13 Public and Private Schools***

Populations with Exclusions or Low Participation	Percentage (Confidence) Intervals of Students At or Above NAGB Achievement Levels		
	Basic (256)	Proficient (294)	Advanced (331)
China In-school Population, Restricted Grades, 20 Provinces & Cities	88.5 - 90.9	43.3 - 47.6	06.7 - 08.9
England Low Participation	65.3 - 70.9	18.2 - 23.0	01.3 - 02.5
Portugal In-school Population, Restricted Grades	50.9 - 55.4	08.2 - 10.6	00.2 - 00.5
São Paulo, Brazil Restricted Grades	28.3 - 33.0	03.3 - 04.8	00.0 - 00.3
Fortaleza, Brazil In-school Population, Restricted Grades	23.6 - 28.0	02.2 - 03.5	00.0 - 00.1
Maputo and Beira, Mozambique In-school Population, Low Participation	16.7 - 20.9	00.5 - 01.2	00.0 - 00.1

FIGURE 5: ***Graphic Illustration of Percentage (Confidence) Intervals At or Above NAGB Achievement Levels for 1991 IAEP Mathematics for Comprehensive Populations -- Age 13 Public and Private Schools***

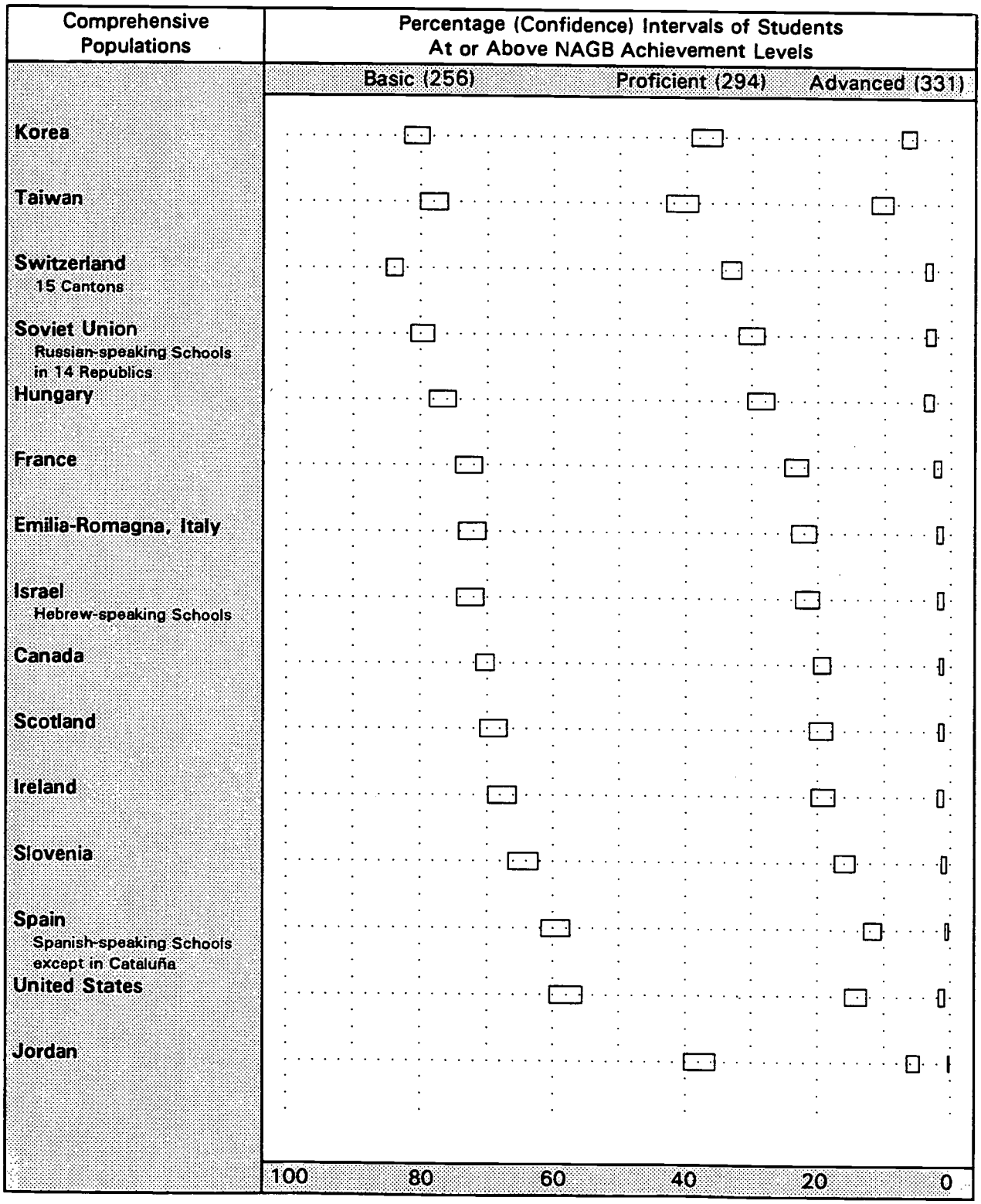


FIGURE 6: ***Graphic Illustration of Percentage (Confidence) Intervals At or Above NAGB Achievement Levels for 1991 IAEP Mathematics for Populations with Exclusions or Low Participation -- Age 13 Public and Private Schools***

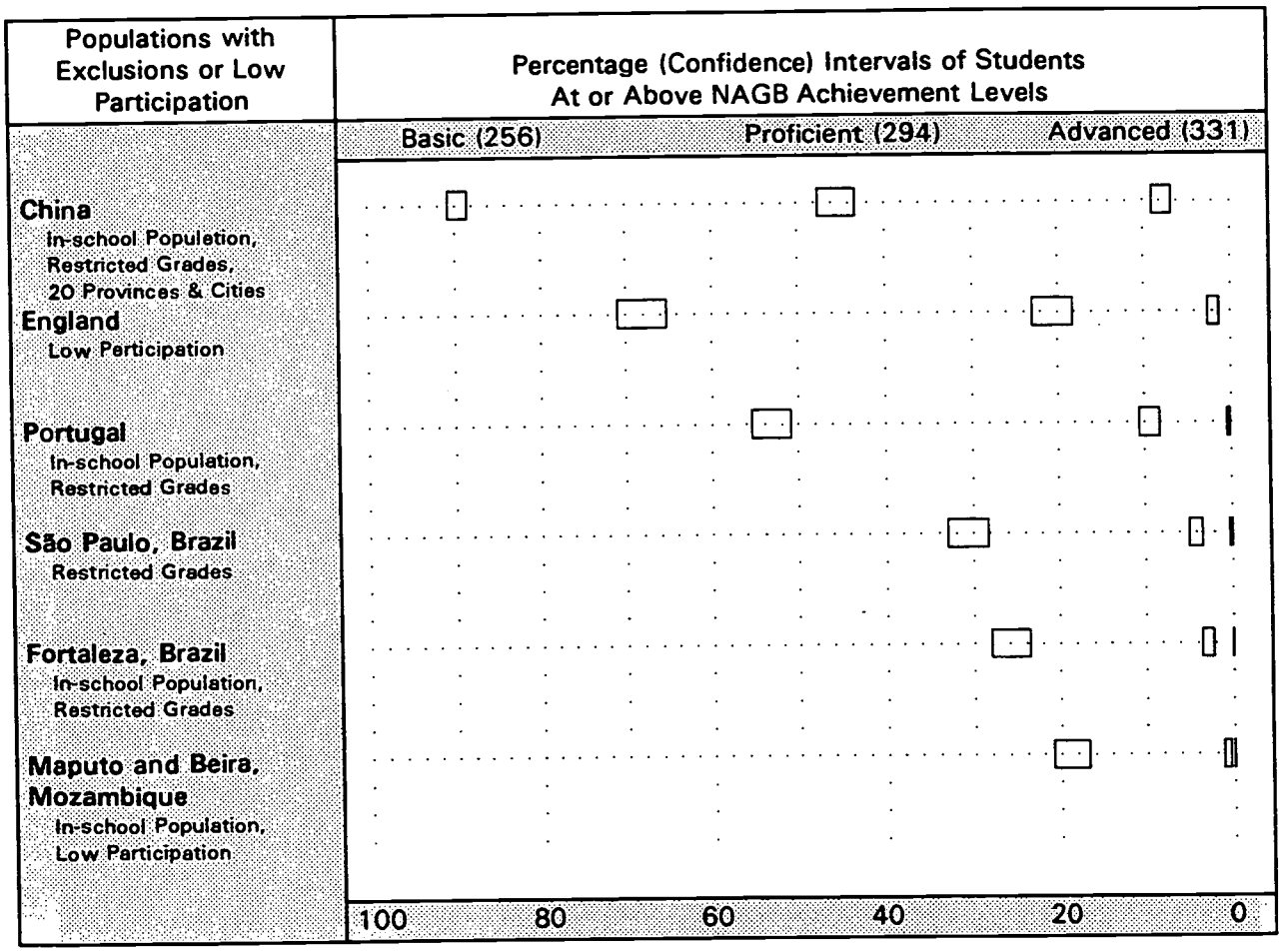


TABLE 6: ***Percentages At or Above NAGB Achievement Levels and Standard Errors for 1990 NAEP Mathematics -- Grade 8 Public Schools***

	Percentage of Students At or Above NAGB Achievement Levels		
	Basic (256)	Proficient (294)	Advanced (331)
NATION	57 (1.4)	19 (1.2)	2 (0.4)
Northeast	65 (3.7)	26 (3.1)	3 (1.0)
Southeast	48 (3.0)	15 (2.2)	2 (0.6)
Central	61 (2.5)	20 (2.1)	2 (0.6)
West	57 (2.6)	19 (2.5)	3 (0.7)
STATES			
Alabama	47 (1.6)	12 (0.8)	1 (0.2)
Arizona	55 (1.8)	16 (1.1)	1 (0.4)
Arkansas	51 (1.3)	12 (1.0)	1 (0.2)
California	51 (1.6)	16 (1.3)	2 (0.4)
Colorado	64 (1.1)	22 (1.0)	2 (0.4)
Connecticut	66 (1.3)	26 (1.1)	4 (0.4)
Delaware	55 (1.3)	19 (0.9)	2 (0.5)
Dist. Columbia	21 (1.0)	4 (0.7)	1 (0.2)
Florida	49 (1.4)	15 (1.0)	2 (0.4)
Georgia	53 (1.5)	17 (1.3)	3 (0.5)
Hawaii	45 (1.0)	14 (0.8)	2 (0.4)
Idaho	70 (1.2)	23 (1.4)	2 (0.4)
Indiana	63 (1.6)	21 (1.2)	3 (0.6)
Iowa	76 (1.1)	30 (1.5)	4 (0.5)
Kentucky	51 (1.8)	14 (0.9)	1 (0.2)
Louisiana	39 (1.7)	8 (1.0)	1 (0.2)
Maine	xxx (xxx)	xxx (xxx)	xxx (xxx)
Maryland	56 (1.7)	20 (1.2)	3 (0.6)
Massachusetts	xxx (xxx)	xxx (xxx)	xxx (xxx)
Michigan	60 (1.4)	20 (1.4)	2 (0.4)
Minnesota	74 (1.3)	29 (1.2)	4 (0.4)
Mississippi	xxx (xxx)	xxx (xxx)	xxx (xxx)
Missouri	xxx (xxx)	xxx (xxx)	xxx (xxx)
Nebraska	74 (1.1)	30 (1.4)	4 (0.6)
New Hampshire	71 (1.6)	25 (1.2)	3 (0.5)
New Jersey	65 (1.6)	25 (1.3)	4 (0.5)
New Mexico	51 (1.3)	13 (0.9)	1 (0.3)
New York	57 (1.7)	19 (1.0)	3 (0.5)
North Carolina	44 (1.4)	11 (0.8)	1 (0.4)
North Dakota	81 (1.6)	34 (2.0)	4 (0.6)
Ohio	60 (1.4)	19 (1.2)	2 (0.3)
Oklahoma	59 (1.6)	17 (1.3)	2 (0.5)
Pennsylvania	63 (2.0)	21 (1.5)	2 (0.4)
Rhode Island	55 (0.9)	18 (1.0)	2 (0.3)
South Carolina	xxx (xxx)	xxx (xxx)	xxx (xxx)
Tennessee	xxx (xxx)	xxx (xxx)	xxx (xxx)
Texas	52 (1.7)	16 (1.0)	2 (0.4)
Utah	xxx (xxx)	xxx (xxx)	xxx (xxx)
Virginia	58 (1.6)	21 (1.6)	4 (0.8)
West Virginia	49 (1.2)	12 (0.9)	1 (0.2)
Wisconsin	72 (1.7)	29 (1.5)	4 (0.5)
Wyoming	71 (1.3)	24 (1.0)	2 (0.3)
TERRITORIES			
Guam	27 (1.0)	5 (0.6)	1 (0.2)
Virgin Islands	10 (1.1)	1 (0.4)	0 (0.1)

*From Mullis *et al*, 1993, Table 4, p. 10.

TABLE 7: ***Percentages At or Above NAGB Achievement Levels and Standard Errors for 1992 NAEP Mathematics -- Grade 8 Public Schools***

	Percentage of Students At or Above NAGB Achievement Levels		
	Basic (256)	Proficient (294)	Advanced (331)
NATION	61 (1.2)	23 (1.1)	3 (0.5)
Northeast	59 (3.9)	25 (3.0)	5 (1.4)
Southeast	53 (1.6)	16 (1.0)	1 (0.4)
Central	70 (2.8)	28 (3.0)	3 (0.7)
West	62 (2.7)	24 (2.1)	4 (1.1)
STATES			
Alabama	44 (2.0)	12 (1.1)	1 (0.3)
Arizona	61 (1.8)>	19 (1.4)	2 (0.4)
Arkansas	50 (1.7)	13 (1.0)	1 (0.3)
California	55 (2.0)	20 (1.4)	3 (0.7)
Colorado	69 (1.3)>	26 (1.3)>	2 (0.5)
Connecticut	69 (1.4)	30 (1.1)>	4 (0.6)
Delaware	57 (1.2)	18 (1.1)	3 (0.4)
Dist. Columbia	26 (1.3)>	6 (1.0)	1 (0.2)
Florida	55 (1.9)	18 (1.3)	2 (0.4)
Georgia	53 (1.5)	16 (1.0)	1 (0.3)
Hawaii	51 (1.2)>>	16 (0.8)	2 (0.4)
Idaho	73 (1.1)	27 (1.2)	3 (0.4)
Indiana	66 (1.5)	24 (1.3)	3 (0.4)
Iowa	81 (1.2)>	37 (1.4)>	5 (0.7)
Kentucky	57 (1.3)>	17 (1.1)	2 (0.4)
Louisiana	42 (2.0)	10 (1.2)	1 (0.2)
Maine	77 (1.3)	31 (1.9)	4 (0.6)
Maryland	59 (1.5)	24 (1.3)	4 (0.6)
Massachusetts	68 (1.5)	28 (1.4)	3 (0.5)
Michigan	63 (1.6)	23 (1.7)	3 (0.5)
Minnesota	79 (1.2)>	37 (1.2)>>	6 (0.7)>
Mississippi	38 (1.5)	8 (0.8)	0 (0.2)
Missouri	68 (1.6)	24 (1.3)	3 (0.4)
Nebraska	75 (1.2)	32 (1.9)	4 (0.5)
New Hampshire	77 (1.0)>	30 (1.5)>	3 (0.6)
New Jersey	67 (1.8)	28 (1.4)	4 (0.6)
New Mexico	54 (1.4)	14 (1.0)	1 (0.3)
New York	62 (2.3)	24 (1.6)>	4 (0.6)
North Carolina	53 (1.5)>>	15 (1.0)>	1 (0.3)
North Dakota	82 (1.3)	36 (1.7)	4 (0.6)
Ohio	64 (2.0)	22 (1.4)	2 (0.5)
Oklahoma	65 (2.0)	21 (1.2)>	2 (0.3)
Pennsylvania	67 (1.7)	26 (1.5)	3 (0.7)
Rhode Island	62 (1.2)>>	20 (1.3)	2 (0.3)
South Carolina	53 (1.2)	18 (1.1)	2 (0.5)
Tennessee	53 (1.8)	15 (1.2)	1 (0.4)
Texas	58 (1.5)>	21 (1.4)>	4 (0.6)
Utah	72 (1.3)	27 (1.1)	3 (0.5)
Virginia	62 (1.6)	23 (1.2)	3 (0.5)
West Virginia	53 (1.5)	13 (0.9)	1 (0.2)
Wisconsin	76 (1.9)	32 (1.4)	4 (0.6)
Wyoming	73 (1.3)	26 (1.0)	2 (0.5)
TERRITORIES			
Guam	30 (1.4)	7 (0.7)	1 (0.2)
Virgin Islands	13 (1.0)	1 (0.3)	0 (0.1)

*From Mullis *et al*, 1993, Table 4, p. 10. >>The value for 1992 was significantly higher than the value for 1990 at about the 95 percent certainty level. <<The value for 1992 was significantly lower than the value for 1990 at about the 95 percent certainty level. These notations indicate statistical significance from a multiple comparison procedure based on the 37 jurisdictions participating in both 1992 and 1990. If looking at only one state, then > and < also indicate differences that are significant. Statistically significant differences between 1990 and 1992 for the state comparison samples for the nation and regions are not indicated. (xxx) Did not participate in the 1990 Trial State Assessment.

To illustrate the contributions made by the four sources of error to the confidence interval calculations, the overall variance estimate related to the three NAGB achievement levels was broken down into components for the U.S. sample. These component estimates, and their percentage of the total variance estimates are given in TABLE 8. Note that the components were derived using Equations 1 through 5, respectively, found in the Methodology section. Also note that the first four components have been averaged over the five sets of plausible values.

TABLE 8: Variance Component Estimates for the 1991 U.S. IAEP Sample

Variance Component	NAGB Achievement Level					
	256		294		331	
	Estimate	% of σ^2	Estimate	% of σ^2	Estimate	% of σ^2
$\bar{\sigma}_r^2$	.000005	4	.000002	3	.00000004	1
$\bar{\sigma}_s^2$	.000062	43	.000026	41	.0000020	47
$\bar{\sigma}_r^2 + \bar{\sigma}_s^2$	.000067	47	.000028	44	.0000021	49
$\bar{\sigma}_r^2 + 2\bar{\sigma}_s^2$	.000129	90	.000054	85	.0000041	95
σ^2	.000142	--	.000063	--	.0000043	--

The largest contributors to the overall variance estimate are those related to sampling issues. First, the sample-to-population error was found to always be quite substantial. Second, this sampling error is then doubled due to the design effect. This does indicate, however, that even better estimates can be realized by collecting more data.

CONCLUSIONS

In an article by Howard Wainer (1993) entitled "How Much More Efficiently Can Humans Run Than Swim?", performances across two distinctly different sports were compared. Wainer concluded that the distance that runners can traverse, in a fixed amount of time, is about 3.75 times that of swimmers. With this information and their own personal performance data, runners could estimate how far they could swim in, say, five minutes, even if they do not know how to swim. Needless to say, these results are of interest to athletes who are intrigued by cross-sport comparisons. Critics might argue that running and swimming are sufficiently dissimilar as to render any link between them meaningless.

Educators and policymakers have also been interested in linking performance as measured by various assessment instruments. If feasible, such links could provide comparisons across disparate populations without the cost of additional testing and in cases where the application of certain assessments is impossible, for example, due to differences in native language. Such linking studies also have their critics who cite various concerns, including problems with differences in underlying cognitive constructs and motivation.

The present study examined a linking of mathematics proficiencies as measured by the NAEP and IAEP assessments. This could be considered an almost "ideal" setting for such an empirical research project. Both assessment instruments were constructed and scored in similar fashion. The testing environments were also quite comparable, and results from a sample of U.S. students, who were evaluated with both instruments, provided a clear picture of the relationship between the IAEP and NAEP assessments. Given this relationship, a linking procedure was easily found.

One goal of this study was the development of methodology for estimating the percentages of IAEP populations falling above cut-scores on the NAEP scale. Associated standard errors were also developed so that the precision of the estimates of percentages could be evaluated. These standard errors accounted for four sources of

uncertainty related to regression estimation, sample-to-population estimation, design effects, and measurement error.

A strong linear relationship between the IAEP and NAEP mathematics assessments was found, based on the 1992 U.S. linking sample. Percentage intervals based on the 1992 U.S. IAEP (linking sample) assessment captured the estimated NAGB achievement level percentages obtained from the 1992 NAEP assessment. Also, a cross-validation study indicated that the proposed methodology was correct and appropriate, at least for these data.

This methodology was then applied to results from the 1991 IAEP assessment. The three NAGB achievement levels were used as cut-score examples. Derived estimates for the 1991 IAEP populations were found to be consistent with other available results. Associated standard errors were found to be influenced most by sample-to-population and design effects.

While this study could be considered a success, some caveats must be kept in mind when considering the results:

- We assumed that the relationship between the IAEP and NAEP assessments observed in the 1992 U.S. linking sample also holds for other countries that were assessed in 1991.
- There were differences in IAEP and NAEP sample definitions, such as type of schools surveyed and age or grade of students.
- Other sources of estimation error, besides the four accounted for, were assumed to be insignificant.

Of course there is no guarantee that the link established in this study will hold for subsequent years. Also, the linking methodology that was developed for these assessments may not perform as well in other less "ideal" situations.

Despite these imperfections, this study has demonstrated that interesting and reasonable empirical linking results can be obtained if equal amounts of appropriate methodology and caution are properly applied.

REFERENCES

Blais, J.-G. (1992). *IAEP Technical Report: Volume Two.* Princeton, NJ: Educational Testing Service, International Assessment of Educational Progress.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443-459.

Johnson, E. G., & Allen, N. L. (1992). *The NAEP 1990 technical report.* (No. 21-TR-20) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Linn, R. L. (in press). Linking results of district assessments. *Applied Measurement in Education.*

Mislevy, R. J. (1985). Estimation of group effects. *Journal of the American Statistical Association, 80,* 993-997.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Mislevy, R. J. (1992). *Linking Educational Assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Bock, R. D. (1989). *PC-BILOG 3: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in National Assessment of Educational Progress. *Journal of Educational Statistics, 17,* 131-151.

Mullis, I.V.S., Dossey, J.A., Owen, E.H., & Phillips, G.W. (1963). NAEP 1992 Mathematics Report Card for the Nation and the States. Washington, D.C.: U.S. Department of Education.

National Assessment Governing Board (1993). Positions on the future of the National Assessment of Educational Progress. Washington, DC: author.

National Council on Education Standards and Testing (1992). *Raising Standards for American Education.* Washington, DC: author.

National Education Goals Panel (1991). *The national education goals report: Building a nation of learners.* Washington, DC: author.

Rogers, A. M. (1991). *NAEP-MGROUP: Enhanced version of Sheehan's software for the estimation of group effects in multivariate models* [Computer program]. Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Wainer, H. (1993). How much more efficiently can humans run than swim? *Chance, 6,* 17-21.

APPENDIX I: DATA COLLECTION

School Sample

The goal for data collection was to administer the IAEP assessment in a sample of 75 NAEP schools and approximately 30 students per school. Eligible schools were those scheduled for at least one grade 8/age 13 mathematics session in the 1992 national (main) NAEP assessment.

In order to secure a school sample, Westat supervisors conducting main NAEP assessments were asked to select one to three eligible schools from among their assignments that might be willing to cooperate. Since it was important to secure an approximately representative sample of schools, supervisors were instructed that the first NAEP school they select had to be an urban public school, the second had to be a public school but could either be urban or rural, and the third school could be a private school. The definition of "urban" was any school located in or around a major metropolitan area. Once a school was identified, supervisors called Westat in order that the school demographics could be checked for "representativeness."

Supervisors re-contacted eligible schools by phone to discuss participation in this project. Subsequently, schools were mailed two informational letters: "Linking NAEP Results to International Comparative Statistics" describing the rationale of the study, and the other a "Dear Principal:" letter describing the project, as well as asking the school for cooperation with the study.

As cooperating schools were identified, the school demographics (e.g., public, private, Catholic, and geographic region) were checked against the percentages sampled for the grade 8/age 13 schools in the NAEP sample for comparability.

Data Collection

National Assessment supervisors carried out data collection during the period of March 16 through April 30, 1992. Since the conduct of the National Assessments were well under way, there was some concern over how the schools would respond to supervisors re-contacting them for cooperation in study. Although this did not turn out to be a problem, if future linkage studies involving NAEP and other assessments are to be conducted, participating schools should be informed during the initial stages of NAEP data collection.

To comply with the project design, IAEP assessments could not be scheduled until after the NAEP mathematics main assessment session(s) in cooperating schools had been completed.

Only those students (eighth-graders or 13-year-olds) who were actually assessed in a main NAEP mathematics spiral session were eligible for this study. (Spiral sessions were used for self-paced administration of the main NAEP mathematics assessment. Other sessions were devoted to audiotape administration of special mathematics tasks.) There were two means by which supervisors could prepare the paper work for student selection and notification. In schools where main NAEP assessments had already been conducted, supervisors were instructed to revisit the school. In schools where main NAEP assessments had not yet been conducted, supervisors were instructed to select students for the IAEP study after conducting the main NAEP mathematics spiral session.

Student Sample

Depending on the number of students actually assessed in a mathematics spiral session, supervisors were instructed to sample approximately 30 students per school as follows:

- If 35 students had been assessed in a school with only one mathematics spiral session, all 35 students were to be invited to the IAEP session.

- If more than 35 students were assessed, every other student up to 30 students would be sampled.

- If two mathematics spiral sessions were conducted in a school, supervisors were to ask the school coordinator which of the two groups of students she or he preferred be invited; if no preference was stated, then every other student up to 30 students would be sampled.

Supervisors discussed with the school coordinator arrangements for an appropriate testing room and the school's preference for notifying teachers and students of the IAEP assessment. If preferred, supervisors prepared and left with the school coordinator notification letters for teachers and NAEP Appointment Cards for students to be distributed a few days before the session. In addition, if the school so requested, parent information letters were also provided. Provisions for student incentives, authorized at one dollar per student, were also discussed at this time.

Preparation of the Administration Schedule

The names, demographic information, and NAEP Booklet Ids of students selected for the IAEP study were recorded onto a new administration schedule. The top section of the administration schedule was completed as it would be for NAEP, except in the space labeled "Session Type" supervisors were to write in "IAEP II." The top of the administration schedule was also to be labeled "IAEP II" so that it could be readily distinguished from that of the NAEP Schedule.

Conducting the Sessions

A copy of Administration Instructions as well as session scripts were mailed to supervisors in advance of their first assessment date. IAEP sessions were conducted in a fashion similar to that of NAEP sessions. Booklet preparation was the same except that the students' NAEP Booklet ID was entered and gridded onto the booklet front cover. Each session required 90 minutes of students' time and was conducted by using a standardized script.

The IAEP assessment was administered in one of two orders, Order 1, and Order 2, as indicated by the script heading. The order was determined by having the supervisors flip a coin: if the coin came up heads the school was assigned "Order 1;" tails, the school was assigned "Order 2." The order number was entered at the top of the Administration Schedule.

Make-up sessions were not required; there were no school, teacher, or excluded-student questionnaires to distribute; and unlike NAEP mathematics spiral assessments, the session did not require the use of calculators.

At the end of the session all assessment materials were accounted for and shipped no later than one day after completion of the session to National Computer Systems for processing.

Results

A total of 74 schools and a sample of approximately 1600 students participated in the study. TABLES I-1 and I-2 reflect the number of participating schools by type of school (e.g., Public, Private, Catholic) as well as geographic region and urbanicity by geographic region. Overall, the school sample closely paralleled that of the National Assessment sample.

The student sample was somewhat less than what we anticipated due to the fact that in most of the cooperating schools, fewer than 30 students were originally assessed in the NAEP mathematics spiral sessions. Taking into account absenteeism, the resulting numbers of students assessed in the IAEP sessions were less than expected. However, it was still adequate for the purposes of determining a linking function.

TABLE I-1: Percentage of Schools by Type and Region

Region	Public	Private	Catholic	Total
Northeast	12	4	5	21 (28)
South	13	0	1	14 (19)
Central	10	3	4	17 (23)
West	15	3	4	22 (30)
Totals	50 (68)	10 (14)	14 (19)	74

TABLE I-2: Percentage of Schools by Urbanicity and Region

	Urban			Rural	Total
Region	*SMSA	**MSA	Total Urban	Non-MSA	
Northeast	15	3	18 (24)	3 (4)	21 (28)
South	0	12	12 (16)	2 (3)	14 (19)
Central	8	3	11 (15)	6 (8)	17 (23)
West	11	4	15 (21)	7 (9)	22 (30)
Totals	34 (46)	22 (30)	56 (76)	18 (24)	74

* Standard Metropolitan Statistical Area

** Metropolitan Standard Area

APPENDIX II: NOTATION

Sample Sizes

n_e: number of examinees in the linking sample (i.e., 1992 US IAEP/NAEP linking sample)

n_a: number of examinees in the application sample (e.g., 1991 Korean IAEP sample)

Subscripts

i: linking sample ($i = e$) or application sample ($i = a$)

j: examinee ($j = 1, 2, ..., n_i$)

k: imputation ($k = 1, 2, ... , 5$)

Scores

$s_{ij(k)}$: imputed proficiency value on IAEP instrument

$x_{ij(k)}$: deviation score on IAEP instrument (i.e., $x_{ij(k)} = s_{ij(k)} - \frac{\sum_{j=1}^{n_e} s_{ej(k)}}{n_e}$)

Note: Deviations are always derived with respect to the linking sample mean.

$y_{ij(k)}$: imputed scaled score on NAEP instrument

Regression Parameters

$\hat{\alpha}_{(k)}$: sample intercept estimate from a linear regression of NAEP scaled scores on IAEP deviation scores (k^{th} imputation)

$\hat{\beta}_{(k)}$: sample slope estimate from a linear regression of NAEP scaled scores on IAEP deviation scores (k^{th} imputation)

$RMSE_{(k)}$: sample root mean square error estimate from a linear regression of NAEP scaled scores on IAEP deviation scores (k^{th} imputation)

Sample Means and Variances

For sample values $w_1, w_2, \ldots, w_n$

$$\text{Ave}(w_j) = w_\cdot = \frac{\sum_{j=1}^{n} w_j}{n} \quad \text{and} \quad \text{Var}(w_j) = \frac{\sum_{j=1}^{n} (w_j - w_\cdot)^2}{n - 1}$$

Miscellaneous

Let c denote a cut-score on the NAEP scale (e.g., 331, 294 or 256), $\Phi[\cdot]$ the standard normal cumulative distribution, and $\phi[\cdot]$ the standard normal density. For convenience, let

$$z_{ij(k)} = \frac{c - \hat{\alpha}_{(k)} - \hat{\beta}_{(k)} x_{ij(k)}}{RMSE_{(k)}}$$

APPENDIX III: COMPUTER PACKAGES

BILOG (Mislevy & Bock, 1983)

This program estimates the parameters of a three-parameter logistic item response model. Marginal maximum likelihood (Bock & Aitkin, 1981) and Bayes marginal modal (Mislevy, 1986) solutions are available.

TBLT

This is an ETS in-house program that performs scale transformations. The program is appropriate for situations in which independent estimates of item parameters, based on independent samples, must be expressed in the same metric. TBLT finds the optimal linear transformation that will minimize the weighted mean squared difference between corresponding test characteristic curves (Stocking & Lord, 1983).

MGROUP (Rogers, 1991)

Given item parameters, this program calculates predictive proficiency distributions conditional on background variables. (The NAEP and IAEP technical manuals can be referenced for a list of the conditioning variables employed.) MGROUP employs a variant of the EM algorithm described in Mislevy (1985).

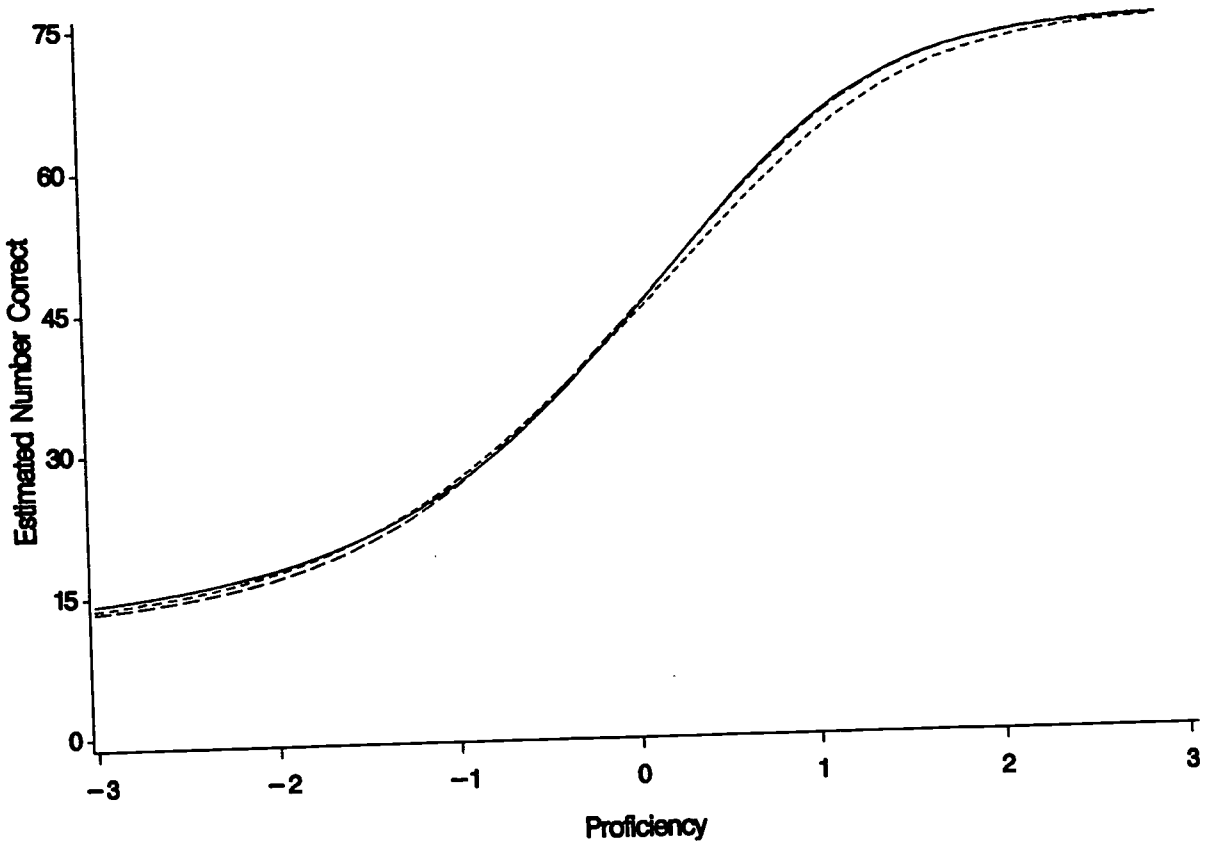
APPENDIX IV: TBLT RESULTS

The 75 items from the IAEP assessment were calibrated twice: in 1991 based on a composite sample consisting of students selected across all participating IAEP populations, and then again in 1992 based on the U.S. linking sample results. As a comparison of the resulting scales, the TBLT procedure (Stocking & Lord, 1983) was applied to the corresponding test characteristic curves.

All 75 IAEP items were used in the analysis. The 1991 IAEP parameter estimates were taken as the "standard" to which the 1992 linking sample parameter estimates were adjusted. The program required four iterations to find an optimal linear function that minimizes a weighted squared difference between the two test characteristic curves. The final slope and intercept estimates were .93 and -.03, respectively.

FIGURE IV-1 displays test characteristic curves based on the 1991 IAEP results, and the original and adjusted 1991 linking sample results. Such relatively close agreement between curves was also evident when the corresponding 75 individual item response functions were compared.

FIGURE IV-1: Test Characteristic Curves for IAEP Mathematics Based on Calibrations for 1991 (solid line), Original 1992 (short dashed line), and Adjusted 1992 (long dashed line)



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").