Feasibility Studies of Two-Stage Testing in Large-Scale Educational Assessment: Implications for NAEP

R. Darrell Bock, *University of Chicago*
Michele F. Zimowski, *National Opinion Research Center*

Commissioned by the NAEP Validity Studies (NVS) Panel
May 1998

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U. S. Department of Education or the American Institutes for Research.

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Patricia Dabbs
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
1791 Arastradero Road
PO Box 1113
Palo Alto, CA 94302
Phone: 650/ 493-3550
Fax: 650/ 858-0958

Contents

Introduction

This report examines the potential of adaptive testing, two–stage testing in particular, for improving the data quality of the National Assessment of Educational Progress (NAEP). Following a discussion of the rationale for adaptive testing in assessment and a review of previous studies of two–stage testing, this report describes a 1993 Ohio field trial of two–stage assessment carried out, under the direction of the authors, by the National Opinion Research Center (NORC). The trial was part of a larger methodological study of science assessment at school–leaving age supported by the National Science Foundation (NSF) and the Office of Educational Research and Improvement (OERI). This report summarizes the instrument design, procedures, and results of the field trial and discusses implications for the conduct of assessment generally, and for NAEP specifically. A technical appendix outlines the measurement justification for two design prototypes and describes procedures used in analyzing the data.

The Place of Adaptive Testing in Educational Assessment

A fundamental result of classical test theory is that, apart from the effects of guessing, dichotomously–scored test items supply the most information for measurement purposes when the probability of the examinee responding correctly is exactly one–half. Adaptive testing procedures use prior information about the examinee to choose test items that satisfy this requirement as closely as possible, while also having a strong relationship to the proficiency being measured. Such procedures require fewer items and less administration time—one–half to one–third or even less—than a conventional test with the same reliability. Different examinees will, of course, be presented by different items, but modern item response theoretic (IRT) methods of analyzing the data make it possible to estimate scores for the examinees comparably on the same proficiency scale.

The most efficient adaptive procedures, those requiring the fewest numbers of items per examinee for a given level of reliability, make use of computerized selection and presentation of items. Computerized adaptive testing (CAT) calculates a provisional value for the proficiency of the examinee after the response to each item, then chooses the next item from a set of highly discriminating items in the system's item pool that have near 50 percent probability of correct response at that value (see Wainer, 1990; Bock and Mislevy, 1982; Owen, 1969). When a large enough item pool is available, CAT applied in an unselected population of examinees can yield efficiencies three or more times greater than a nonadaptive test constructed from the same pool—that is, CAT can attain test reliabilities equal to that of a conventional test with three or more times as many items. The technology of CAT is now well developed: it is routinely available to college students taking the College Board Graduate Record Examination, it is presently being implemented by the Department of Defense for accessions testing with the Armed Services Vocational Aptitude Battery, and is gaining use in business and industry for personnel selection.

As computers become widely available in schools, the prospect of administering NAEP by CAT becomes very attractive. In addition to reduced testing time, many benefits to data quality would result. For example:

1. *More flexible scheduling of testing*: individual students can begin and end the CAT session at any time.
2. *More varied item displays*: dynamic diagrams, photographs, voice and sound via earphones, etc.
3. *Less reliance on multiple–choice items*: the student's keyword response can be a number, word, or several words that the computer's scoring protocol will classify as correct or incorrect.
4. *Suppression of omitted items*: a valid response is required for the next item presentation and random responding is reduced because the difficulties of the presented items are adjusted to the student's capacity.
5. *Better measurement precision*, especially in the tails of the proficiency distribution in the population of students.
6. *Better motivation during the testing session*: adaptive testing avoids presenting discouragingly difficult items to the examinee or items so easy as to make the test seem "dumbed down."
7. *Elimination of present NAEP procedures of conditioning on student background characteristics in order to strengthen estimation of plausible scores*: the conditioning on provisional estimates of student proficiency within the adaptive session is much stronger for this purpose than conditioning on background characteristics.

Computerized adaptive testing, however, has some disadvantages. It precludes the use of extended responses to problem solving exercises, essay questions, etc., which cannot be scored on–line by the computer. These types of exercises would have to be administered nonadaptively, perhaps in paper–and–pencil mode if diagrams, constructions, or calculations are involved that cannot be handled on the screen. An equally serious problem for CAT in assessment is the long development time required to create an effective and robustly operating computerized test. An item pool containing a large number of well–tested items is required, and the work of creating such pools will be heavy when the range of proficiency to be covered is large and more than one subject matter area is to be tested. To avoid the cost of preparing computer presentations of items that are later rejected, much of the item field testing must be carried out in paper–and–pencil mode. Further studies must then be conducted to adjust for presentation–modality effects. Until computers in the school become more standardized, especially in their keyboards and size and resolution of displays, similar studies may be necessary to allow for equipment differences. For these reasons, group administration of two–stage adaptive tests in paper–and–pencil mode could play an important role in the transition to a fully developed computerized system. It might also have a continuing place in occasional assessments of specialized subjects where the development costs of CAT could not be justified.

In two–stage paper–and–pencil testing, a highly discriminating first–stage test is administered to the examinee in order to classify him or her in one of several broad levels of proficiency in the subject matter. Subsequently, a second–stage test with items

optimized for measurement at the corresponding assigned level is administered to that examinee. Information in item responses at both stages is then combined to obtain a final best estimate of his or her proficiency. This method of testing can, on average, estimate IRT scale scores with measurement precision equal to that of a conventional test containing somewhere between two and three times as many items as the combined first– and second–stage tests. Although not as efficient as fully adaptive computerized testing, it is readily adaptable to present NAEP test administration procedures. It would have some but not all of the advantages of CAT. Features 1 though 4 of the above list would be lost, but 5, 6, and 7 would still be available in two–stage testing.

Two–stage also offers more flexibility in presenting constructed–response exercises in the second–stage. To respond to an essay prompt in CAT, the examinee would have to have adequate typing skills, or, to construct figures and diagrams, skills in using the pointing device. For the most part, constructed–response exercises in CAT would be suitable only for the older age groups tested by NAEP, whereas paper–and–pencil exercises present no special problems for any of the groups. In either response mode the artifact, whether a computer file or a written page, has to be read and rated at a later time, as in current NAEP operations. Once the ratings are available, IRT scoring procedures can utilize optimally the greater amount of information that graded scoring conveys compared to correct–incorrect scoring.

In theory, since the graded rating scale categories extend over a range of proficiency levels, graded scoring should reduce the need for adaptive testing. In practice, however, presenting a task too difficult for the student often results in no response at all or an off–topic response. For this reason, open–ended exercises, such as annotated multiple–step exercises or essay questions, require pretesting for productivity in the target population of test takers just as multiple–choice items require pretesting for difficulty. Based on the pretest results, two–stage testing can then present examinees with exercises that are likely to be response–productive.

An issue to be resolved in a NAEP application of two–stage testing is how the test administration would be carried out. The feasibility studies discussed in this report employed first– and second–stage test administrations separated by an interval of a number of weeks to allow time for scoring the first stage and assigning examinees to second–stage forms. Because it would require the testing teams to visit each school twice, that approach would increase appreciably NAEP field costs. If improved data quality and saving of processing time did not justify the increase, some way to conduct both stages of testing on the same day would have to be found.

A possible solution is to have the teams administer the first–stage test during a morning hour, then score the tests and assign the second–stage forms in time for a second hour of testing in the afternoon. Assuming that the first–stage test contains only multiple–choice items, which is typically the case, and considering that the number of students tested for NAEP in each school is relatively small, the highly portable equipment that is now available for scanning documents and computing test scores should make this approach feasible. Only a notebook computer, with an attached scanner and portable inkjet printer would be required. Special first–stage forms could be prepared with detachable pages from which the scanner could read an optical character ID number and detect the presence of marks in the answer spaces. The computer driving the scanner would then immediately compute IRT scale scores for each examinee, determine the appropriate second–stage form, and print a list of ID numbers and assignments. The entire procedure should not require more than one hour's work on the part of the testing team.

The Case for Credible Student–Level Scores in NAEP

Adaptive testing of achievement is geared to producing scores for individual students. It is motivated and justified by its power to evaluate the examinee's performance using fewer numbers of items than conventional tests, but with comparable reliability. Adopting CAT or two–stage testing would give NAEP the capability of producing credible student–level scores without great increase in testing time or cost. It would, however, significantly change the direction of the assessment as originally conceived by the ECAPE (Exploratory Committee on Assessing the Progress of Education). NAEP was planned as a sample survey of average achievement levels among children to be reported only at the level of large national regions and demographic groups. Facing opposition from education associations and the Congressional leadership to any form of a national test, the committee members specifically excluded any use of the data that would identify students, schools, communities, or states. Paradoxically, this turned out to be advantageous from a measurement point of view because it permitted the test data to be collected by matrix sampling—that is, by sampling students within schools and administering to different students different small samples of test items drawn from much larger sets representing the subject matter domains. When the results from the brief tests are aggregated for large groups of respondents, the large numbers of items represented in the assessment instrument gives the statistical summaries at the group level a high degree of stability and generalizability (see Lord and Novick, 1968, p. 252 ff.).

An indication of the gain in stability at the group level resulting from item sampling is shown in Table 1, adapted from Table 1.1 in Bock and Zimowski (1989). Based on school–level scores from the California Assessment Program, the table shows correlations between sixth–grade average reading scores in two successive years from all public schools having 200 or more sixth–grade students. The assessment instrument consisted of 30 randomly parallel forms containing a total of 420 reading items. The correlations in the table were computed from number–correct scores for matrix samples of 50, 100, and 200 students and 85, 128, and 400 items. Apparent in the table are substantial increases in year–to–year stability of the school scores with increasing numbers of items in the instrument, as well as similar increases with the numbers of students sampled per school. These coefficients are, of course, only lower bounds on the true generalizability of the school scores: they are attenuated by the real variability in the standings of the schools from one year to another. The gains with increased sample sizes are, however, accurately reflected. They corroborate results of matrix sampling theory showing that large samples of items in the assessment instrument are just as important for data quality as large numbers of respondents surveyed.

Table 1— Effect of sampling students and items on year-to-year correlations of sixth-grade mean reading scores in California schools

Number of Students Sampled per Grade	Number of Items in Matrix Sample		
	85	128	400
50	.59	.73	.79
100	.67	.78	.88
200	.76	.81	.93

Because year–to–year changes in national and state mean scores are small, rigorous stability at high levels of aggregation is essential for statistics used in analyzing assessment trends. Error variation arising from both the sampling of items and the sampling of students must be even smaller for stable trends to appear in the results. This means that the adaptive assessment instrument must attain at least the level of generalizability of the present NAEP matrix sampling design. In CAT, this requirement is met automatically by the large number of items necessary for sequential selection of items at many different levels of difficulty. In two–stage testing, it requires the construction of a number of stratified randomly parallel forms of the two–stage test, which are then assigned randomly to the students selected for testing in each school. These forms must contain at least as many distinct items as the current NAEP instrument. The two–stage designs discussed in the present report provide for such multiple forms.

At the present time, NAEP continues to rely on matrix sampling without reportable student–level scores despite enabling legislation that now permits state–level reporting and prevailing sentiment in education favorable to reporting at the student level. The lack of scores for individual students has persistently raised concerns about the validity of NAEP results. Most prominent is the question of whether students have any motivation to perform well on the NAEP tests when neither they nor their parents will receive any report of their test scores. The students know only that they are selected at random to take the tests, that the tests are not directly related to their studies, and that they will hear nothing further of the results. The potential effects of testing under these conditions are troublesome, not only because they may be depressing performance levels nationally, but also because they may affect various demographic groups differentially. Major reporting categories of NAEP such as sex, SES, and age, may be confounded with effects of motivation. At certain ages, for example, boys may be less motivated than girls, or low SES groups may be less motivated than high SES groups. Similarly, older and more test–wise students may be less motivated than younger students, in which case interpretation of gains across school grades would be compromised.

Little is known objectively about the presence or extent of motivational effects in the NAEP data. Kiplinger and Linn (1996) embedded NAEP items in booklets of the Georgia State Assessment program and compared their percent correct statistics with those of the same items in the Georgia State NAEP trial. Differences were small and gave no clear evidence of effects that could be attributed to differences in

motivation. This lack of positive results does not bear directly on the question of motivational effects of student–level score reporting, however; at the time of the study the Georgia Assessment reported only at the school level and above.

In a study of effects of extrinsic motivation, O'Neill, Sugrue, and Baker (1996) paid eighth– and twelfth–grade students one dollar for each correct response on 41 and 44 item tests, respectively. Numbers of correct responses were compared with those of three control groups who were offered only non–monetary incentives. Statistically significant differences in favor of the monetary incentive was observed for the eighth graders, but the mean score was only 2.6 percent higher than that of the control groups. No significant difference was found for twelfth graders. A question unanswered by the study is the effect of the type of test items. If the items require knowledge of facts or procedures that the students do not know, mere eagerness to succeed will not help. In contrast, if the task was, for example, to write an extended response to an essay topic, it seems safe to assume that promised payment by the word would have a positive effect on production. The implications of this study for the NAEP motivation question are unclear.

The absence of any returned information almost certainly makes recruitment of schools for the state and national samples more difficult as well. School officials must agree to cooperate knowing that their students will gain little, if anything, from the lost classroom time and that parents will have little or no interest in the activity. They can justify the time and attention devoted to NAEP testing only on tenuous grounds of future progress in education for the state as a whole. In contrast, computerized reports to students and parents showing scores in relation to state or national norms, with some explanation of what the tests measure, would make the assessment more rewarding for those participating at the local level and create a more favorable attitude on the part of the principals, superintendents, and school board members who must accept participation in the NAEP testing.

The present enabling legislation for NAEP requires that personal identifiable information remain confidential. If reports to students mailed first class to their home address are considered confidential, as they are in many business and professional matters, then the legislation does not preclude such reports. If they are not so considered, then a change in the present legislation would be necessary. In either case, a change would be required in NAEP's present policy of identifying student records only by code numbers on examinee rosters that are kept by the participating schools. NAEP would need records of examinee addresses in order to send reports to parents. Since similar identification of cases and use of addresses is routine in the National Educational Longitudinal Studies, no new precedent is involved.

Background of Two–stage Testing

All present work on adaptive testing, including two–stage, is based on IRT. The so–called "item invariant" property of IRT makes possible the estimation of comparable scores from arbitrary subsets of items measuring the same proficiency, a property not shared by conventional percent–correct scores. IRT scoring makes use of statistical models that account for differences in the difficulty and discriminating powers of the test items or exercises, and the effects of guessing on multiple–choice items.

These models are available for ratings of performance exercises as well as for right–wrong scores of multiple–choice or short–answer questions.

For use in IRT scoring, the items must be previously "calibrated" by estimating parameters of the models from responses of a sample of examinees in the population of potential test–takers. In ongoing assessment programs, these calibrations can be carried out with item response data obtained during operational testing. That is, the calibrating information on exercises for future use can be obtained by including them in the test booklets of the current assessment as so–called "variant" items—items to which examinees will respond, but which will not be used in computing scores in that assessment (see Zimowski, Muraki, Mislevy and Bock, 1995). This calibrating information is especially important in adaptive testing, where the difficulty and discriminating power of the items must be accurately known during instrument development.

Studies of two–stage testing were undertaken, however, before IRT methods became widely available. In the most extensive of such studies, Linn, Rock and Cleary (1969) examined simulated adaptive testing procedures, including two–stage, using data from national administrations of the SCAT and STEP tests. Using responses of 4,885 eleventh–grade students to 190 items covering verbal aptitude, reading achievement, and writing skills, they constructed adaptive and conventional tests from subsets of the items. The authors had a rare opportunity to evaluate the actual predictive powers of the procedures with scores available for approximately two–thirds of these students on the PSAT and College Board Achievement tests administered a year–and–a–half later.

Of the several methods used by the authors to create the first–stage test, the one most similar to the IRT methods discussed in the present paper involved 1) using the number–correct score on the 190–item test to divide the sample into four ordered groups of approximately the same size, 2) in the top and bottom group, computing percent–correct values (*p*–values) for each item, and 3) choosing the twenty items with the largest difference in *p*–values between the groups to make up the first–stage test. Cutting points on this first–stage test then assigned the sample cases to ordered second–stage groups of approximately equal size.

To create the second–stage tests, the authors computed biserial correlations between each of the 190 items and the number–correct scores on the 190 items for cases in each of the second–stage groups. For the test at each of the four levels, they chose, without replacement and excluding items already used in the first–stage test, 20 items with the highest biserial correlation with scores on the 190–item test.

Not having available the item–invariant scoring procedures of IRT, they calculated the second–stage score by fitting least–squares regression equations for predicting the 190–item number–correct scores from the 20–item second–stage number–correct scores. These equation were fitted 1) separately in each of the four second–stage groups, and, to possibly improve the stability of prediction, 2) with a pooled estimate of the common slope coefficient for the groups.

The predictive validities of the second–stage test scores computed in this way were evaluated by correlating them with the scores on the independently administered PSAT Verbal and Math tests, and College Board History and English tests. Similar validities were computed for conventional tests consisting of the 10, 20, 30, 40, and 50 items with the highest point biserial correlations to the total score on the 190–item tests. Among the other adaptive procedures that the authors compared (which are now obsolete), the two–stage procedure showed some of the largest increases in correlations with the external criterion tests relative to a conventional test of the same length

(40 items in the combined first and second stages). As an overall index of the gain efficiency with the two–stage test, they computed the ratio of the number of items on the two–stage test to the average number of items on the conventional test that would be required to obtain the same level of validity. These indices were 3.36 based on separate regression equations for each of the four second–stage groups, and 2.33 based on the regression equation with a common slope.

The results of the Linn, Rock and Cleary study were very favorable to two–stage testing and no doubt inspired Lord's (1971) theoretical study of the topic based on IRT principles. Lord points out that the IRT approach provides direct estimation of the scale scores of the examinees from combined stage–one and stage–two item responses; it also evaluates the relative efficiency of a two–stage test at every point on the score continuum rather than estimating just the average reliability or validity. This property is essential in adaptive testing because the largest gains in measurement precision occur at scale values away from the population mean.

For present purposes, the interesting results in Lord's analysis are that the gain in efficiency between a three– and four–level two–stage test is relatively small, and that including probabilities of chance success in the IRT models degrades the efficiency of two–stage testing considerably, especially when the first–stage test is relatively short. Regrettably, results in Lord's (1971) paper on the effects of chance successes are limited to six–level second–stage tests; they are difficult to compare either with the Linn, Rock and Cleary four–level study or with our investigations of three–level tests. They also assumed many more items (a total of 60 in the two stages) than would be practical in large–scale testing.

To apply Lord's efficiency analysis to testing conditions typical of NAEP, we assumed a two–stage design more similar to that of Linn, Rock and Cleary—namely, 15 or 16 items per subject–matter at each stage. This would easily allow testing of two subjects in a 50–minute period if multiple–choice or short answer items are assumed. In the appendix to the present report, we examine the theoretical efficiencies of two types of designs with these number of items and simulate their application in a manner similar to that of Linn, Rock and Cleary. We also consider the question of how IRT item calibrations for two–stage tests can be carried out in operational assessment data, rather than in data from previous field trials as is usually required in adaptive testing. Because of the large numbers of items needed to insure stability at high levels of data aggregation, as discussed above, calibration in the operational data is essential for NAEP or any similar large–scale assessment program.

The Lord and the Linn, Rock and Cleary studies were carried out in the context of scholastic aptitude testing and do not address many issues of adaptive testing in an assessment environment. More recently, two assessment–oriented studies of two–stage testing have been reported by Bock and Zimowski (1989). They describe an eight–form, two–stage assessment instrument in eighth–grade mathematics evaluated in Illinois and California public schools. The first study was carried out in 32 Illinois schools; after revision of the instrument, the study was repeated in 32 California schools. The trials for both studies were conducted in the field by NORC using procedures simulating an operational assessment.

Teachers participating in these studies administered the first– and second–stage tests on two consecutive class days. During first–stage testing, the students wrote their names and marked their last names and initials into the grid of a machine–scorable answer sheet; they also marked in their gender, date of birth, and a school and teacher

code that the teacher wrote on the blackboard. After responding to the first–stage test in the allowed testing time, the students placed the answer sheet in their test booklets and returned them to the teachers.

Before the administration of the second–stage tests, the teachers scored the 15–item first–stage tests and assigned each student to a second–stage booklet according to the range of number–correct scores on the first–stage test set for each level of the second–stage test; these number–correct ranges were printed on the scoring stencil. After the teachers scored a given student's answer sheet, they selected a consecutive second–stage test booklet at the appropriate difficulty level, and marked the booklet's serial number (which incorporated the form code) into a grid on that student's answer sheet. To guard against the possibility of an omitted form code, we asked that the answer sheets be returned to NORC in the corresponding booklet after the second–stage testing.

Although the teachers reported very few difficulties with the two–stage procedure, many found that the second–stage test was too long to be administered in one class period. A number also complained that scoring the first–stage test and coding information onto the students' answer sheets was too time consuming. In two of the schools, the principal asked NORC personnel to score the pretest and assign the booklets in order to make the testing less demanding of teachers' time.

The IRT item estimation and efficiency analysis in these studies demonstrated the theoretically expected efficiency gains of two–stage testing, and clearly encouraged further studies with other subject matter and grade levels. However, test administration with separate answer sheets is no longer considered suitable for educational assessment, especially at younger ages, and teacher–scoring of the first–stage tests in these studies proved too cumbersome for large–scale practical application. This led us to consider a system of two–stage testing in which all phases of scoring and forms assignment were performed centrally.

The Secondary School Science Assessment Study (SSAS)

As part of a demonstration project in performance assessment of secondary school science learning, we had an opportunity to carry out a much improved approach to two–stage assessment in the state of Ohio. The improvements included expendable test booklets, computer controlled laser printing of personalized second–stage test booklets, and central processing of responses to both first– and second–stage tests. Support for the project, which also included extended–response essay questions and hands–on laboratory exercises, came from the National Science Foundation and the UCLA Center for Research, Evaluation, and Student Testing (CRESST) funded by OERI. A report of the full project is in preparation (see Bock, Doran and Zimowski, 1998). The two–stage study, which was limited to multiple–choice items, encompassed the main subject–matter areas of contemporary secondary school science—Earth sciences, biology, chemistry, and physics.

As an exercise in adaptive testing, the study was especially demanding because of the wide differences in science course work typical of students at school–leaving age. To overcome this difficulty was the main motivation for two–stage study: the objective was to depict outcomes of secondary school science instruction at all levels of science preparation in the complete cohort of students.

The sample

Schools participating in the project were recruited through the offices of the Superintendent of Education of the state of Ohio. To select the school sample, we made use of information from Quality Education Data, Inc., to classify all academic high schools in the state by 1) size of senior class (at or below 165, above 165), 2) urbanicity (urban, suburban, rural), and 3) poverty level (Orshansky percentile at or below 9, above 9). Three schools were selected randomly for each of the 12 cells of this cross–classification. An additional four schools were selected randomly from a list of Ohio vocational high schools. These schools were solicited for participation in the study by mail, including a letter from the state Superintendent requesting the cooperation of districts and schools. After follow–up phone calls by NORC field staff, 33 of the school principals agreed to participate. Additional schools were randomly selected from the same categories as replacements for the three refusals.

One additional school not included in the main study was recruited as a trial site for the field procedures.

The two–stage assessment instrument

The instrument developed for the two–stage study consisted of a student background questionnaire, a 24–item first–stage test containing six items in each science area, and six forms of a three–level second–stage test each with 64–items per level. The test was designed to measure the student proficiencies in science near the end of their secondary school program. The proficiencies general to the four science areas were categorized as follows:

1. Knowledge of scientific terminology and findings.
2. Knowledge of scientific methods and procedures.
3. Understanding of scientific concepts and principles.
4. Proficiency in problem solving.

Sampling of content within each area was stratified in the following subcategories:

1. *Earth sciences*: land, air, water, space.
2. *Biology*: cellular biology, organismic biology, reproduction and genetics, biological diversity.
3. *Chemistry*: the atomic model, states of matter, chemical reactions, quantitative chemistry.
4. *Physics*: mechanics, electricity and magnetism, heat and kinetic theory, sound and optics.

The item pool

Beginning in the fall of 1989, materials for the SSAS were solicited from state, provincial, and national assessment programs that made retired items available for distribution. The generous response of these organizations resulted in a database of over 11,000 multiple–choice items, in some cases with content and cognitive process classifications and item statistics. As they arrived, these items were marked with document and item serial numbers so that they could be conveniently assigned to the

content and proficiency categories of the present study. If subsequently chosen for the assessment instrument, they were entered into a computer data base system by scanning from the original document.

The classification of items by content subcategories and proficiency categories was carried out by persons familiar with the subject matter. If an item did not include a *p*–value from a secondary–school student population, the classifier made a judgment of difficulty by comparing it with similar items of known percent correct. Despite the large size of the resulting item pool, items at suitable levels of difficulty to make up all 720 items required for the six test forms were not found in the collected set. Approximately 60 additional items had to be written especially for the study.

The first–stage test

In constructing the first–stage test, we proceeded on the assumption that a distinct subtest would be required for each science area. With students having backgrounds of different numbers and kinds of science courses, it was unlikely that one could identify a general science proficiency that would be measurable by single test. As specified in the two–stage assessment test prototype described in the appendix, we would therefore have preferred a 64–item first–stage test, with 16–item subtests in the four sciences. Unfortunately, we were offered only two–class periods for paper–and–pencil instruments, one for the first–stage test and the other for the second–stage test. Specifically, we were limited to 45 minutes of test administration time in each period. Although an option might have been to reduce the number of items per subtest to 10, we elected to use a still shorter first–stage test with 6–items per subtest and supplement it with a student background questionnaire asking for students' science and math course–taking history and other activities in science.

Items for the first–stage test were taken from the California Assessment Program twelfth–grade science assessment forms. These items covered all four science areas, and they came with item *p*–values and biserial correlations with total test score based on a large representative sample of California twelfth–grade students. Within each science area, six items having reasonably high biserial correlations and a spread of *p*–values were chosen in each science area for the pretest. In the test booklet, the 24–items were arranged in a random spiral (i.e., in successive sets of four items, each covering the subject areas in random order, and increasing in difficulty from set to set).

The second–stage test forms

The 64–item second–stage instrument assumed 90 minutes of testing time in actual operational use. Even with fully adaptive testing, it would be difficult to measure individual student proficiency accurately in four science areas in less time. With only one class period available to us in the Ohio study, however, and not wishing to test in only two areas or test different students in different areas, we chose to reduce the number of items per form to 32 by removing two items at random from each of the content–by–proficiency subcategories. Since the scores would be reported graphically as a profile of IRT scale scores with standard error bands, we felt that the reduction in score reliability could be tolerated in a field trial. When evaluating the test information functions, we could extrapolate the results from the version with eight items per subtest to a version with 16 items (see the discussion of test information curves below).

The two–stage testing procedure

Administration of the background questionnaire and pretest occurred in January and February of 1991; second–stage testing followed in late April and early May. All procedures for the field study were tried out in the trial school before their use in the main study. All aspects of conducting the study in the field were the responsibility of NORC Operations staff and field representatives.

Principals of the participating schools nominated a teacher, usually a science teacher, as coordinator of testing for the study. NORC Operations obtained from the coordinator a count of the number of senior students eligible for the study (special education students were excluded). The pretest was administered in a class required of all seniors. The teacher of that class served as test administrator. A packet containing instructions for the coordinator and test administrator was sent, along with sufficient numbers of test booklets, to each school via United Parcel Service (UPS). A NORC representative then visited the school to discuss procedures for test administration and immediate return of completed booklets to NORC via prepaid UPS. Booklets from any make–up testing of previously absent students were returned in post–paid business reply envelopes.

Processing of the returned first–stage booklets was carried out by SSAS staff supervised by Dr. Zimowski. Key–entry of student names, ID numbers, questionnaire and item responses was contracted to a commercial service. The resulting data were entered into the database of a large–capacity desktop computing system. For purposes of assigning students to second–stage forms, number–correct scores for the six items in each science area were merged with the questionnaire data.

A computer program was written to assign students to second–stage levels in each area using the following information: number–correct score on the corresponding first–stage test, number of courses in the area, total number–correct score on the first–stage test. With variation to account for different numbers of courses available in the four science areas, the assignment rules were essentially as follows:

1. If area score is 0, 1, or 2, assign to level 1; but if the score is 2 and more than 1 course and total score is 8 or more, assign to level 2.
2. If area score is 3 or 4, assign to level 2; but if the score is 4 and more than 1 course and total score is 12 or more, assign to level 3.
3. If area score is 5 or 6, assign to level 3; but if the score is 5 and no course, assign to level 2.

A personalized second–stage test form for each student who completed the first–stage test was produced by a mainframe–computer–controlled laser printer; it also inserted a colored cover sheet and printed the student's name, then stapled each booklet. Each science area was represented by a contiguous set of eight items, but the order in which the areas appeared rotated from one booklet to another. The items in each area depended on the second–stage level to which the individual student was assigned; the form selection within level was random.

Files controlling the booklet generation were created by a desktop computer program called TEST–BUILDER developed by Dr. Zimowski. The program could input items from many sources, store them in a richly cross–indexed database, and draw them from the database for typesetting and formatting camera–ready copy. Once the item text and graphics have been scanned, edited, and entered into the database, a device–independent file or camera–ready copy for any combination of items could be

prepared in a matter of minutes with a minimum of human intervention. No manual editing, cutting, or pasting of figures was required. The program's expert system automatically placed the pictures in the typeset output and optimally arranged the items and pictures on the pages of the instrument. The booklets were expendable and did not require a separate answer sheet.

In the present study, the page files prepared in this way were transferred to a second device–dependent computer program for laser printing. This program stored the pages of the instrument as digitized images, selected and arranged them in the desired order, overlaid them with typeset information including the student name, and directed the result to the high–speed printer, in this instance a Xerox 9790. Production of each booklet required 20 seconds at a cost of about 45 cents in processing charges. A total of 6,675 booklets were generated in this way.

Because the second–stage testing was limited to one class period in this study, the test booklets consisted of the 32–item short forms. To facilitate orderly distribution to students, booklets were generated and packed for distribution in alphabetical order by student's last name within the classes in which the pretest was administered. The school coordinator was requested to have the booklets administered in the same classes as the pretest. Each school was visited prior to the second–stage testing to discuss administration procedures.

On the last page of each booklet, students were asked to give an address if they wished to have their test results sent to them during the summer; most students did so. Results were reported to the students in the form of a profile of IRT scale scores with mean 250 and standard deviation 50 in the sample. Standard errors and corresponding sample percentile ranks were included.

Data completeness

Thirty–nine of the 40 schools completed the testing and returned the test booklets and make–up booklets. One of the academic high schools did not administer the second–stage tests because of objections of the teaching staff to any outside testing. Another school was non–compliant to the extent of returning only 19 percent of the second–stage booklets. Apart from the lack of cooperation of these schools, the two–stage testing was uneventful and executed responsibly on schedule. Conducted as described above, there was no indication that this form of testing presented any special difficulties for administration by local school personnel.

As a hedge against any unforeseen operational complications, the study allowed approximately six to eight weeks between the two testing stages. Some loss of cases has to be expected with an interval of this length between test and retest, and this was true in the present study. Of the 6,675 students who completed the questionnaire and pretest, 5,375 completed the second–stage tests—a loss of 1,300 cases. Of these, 519 are accounted for by the two non–cooperating schools, leaving 781, or 11.7 percent of the original sample, scattered among the remaining 38 schools. Most of these cases presumably represent students who were not in school on the day of the second–stage test, or who dropped out of school or transferred in their senior year. The number of students responding in any one science area was further reduced because, to accommodate open–ended problem solving exercises, one third of the test booklets included multiple–choice items in only two of the four areas. This brought the number of cases entering the data analysis in each science area to approximately 4,400.

Other well–known types of data loss in testing, such as failure to complete the test or multiple marking of response alternatives, were minimal. We assume that the personalization of the forms with the student's name and the fact that scores would be reported to the student improved the quality of responding.

IRT scaling of the second–stage test forms

For purposes of IRT scaling of items in the easy, intermediate, and difficult booklets of each test form, the booklets included a number of linking items between levels. The original plan was to have four common items between levels in the long form and two in the short form. Thus, each subject–matter area in the three (easy, intermediate, and difficult) test booklets of the long form would contain 48 – 8 = 40 distinct items and of the short form, 24 – 4 = 20 distinct items. An attempt was made to order the items by difficulty and to choose link items near the boundaries between levels.

In the booklets actually constructed, linking did not adhere exactly to this plan because half the items were excluded from the short forms and, among the remaining options, the item choices were constrained by the need to preserve a balance of content and proficiency categories within subject–matter areas in each test booklet. In some test forms, only one item appeared in the link between levels. This was not a cause for concern, given the type of second–stage IRT analysis described in the appendix. When applied simultaneously to all six randomly parallel forms, the analysis aggregates the links across forms to estimate a common latent distribution; thus, even with only a single link item in each form, there would be six items to contribute to linking between levels. Moreover, in this type of analysis the six first–stage items in a given subject area are calibrated jointly with those of the second–stage in the IRT scale construction. They contribute in this instance three additional links between the easy and intermediate levels, and three other links between the intermediate and difficult levels.

As in other IRT applications, a necessary preliminary is the choice of item response model, in this case the one–, two– or three–parameter logistic model for dichotomously scored responses. Because only the pretest items were selected with prior information about their discriminating powers, there is no basis for choosing the one–parameter model. The customary model for multiple–choice items in this type of application is the three–parameter logistic, which insofar as possible corrects for the average effect of chance successes in marking the correct alternative. Considering Lord's (1971) demonstration of loss of information in two–stage testing due to chance successes, however, we attempted to reduce their effect by instructing students taking the first– and second–stage tests as follows:

> *Make the best choice you can, but do not make blind guesses.*
>
> *If you are sure that you do not know the answer to a question, go on to the next question without marking the answer.*

Omitted items were then scored incorrect. There are arguments pro and con for this type of test instruction. The arguments in favor are the loss of information and the technical difficulty of estimating guessing effects accurately, especially for easy items where relatively little guessing occurs. The arguments against are that 1) students who ignore the instruction gain some advantage in score level when the two–parameter

model is assumed, 2) there may be information in wrong responses that can be recovered with the multiple nominal categories model (see Bock, 1997), and 3) these instructions would be inconsistent with computerized adaptive testing in which the examinee is required to respond to the current item before the next item will be presented. We find the latter argument persuasive, but we included the above instructions for the sake of investigating their effects on item parameter estimation and information yield in scoring. When item parameters are estimated by the maximum marginal likelihood method (Bock and Aitkin, 1981), a likelihood ratio statistic is available for testing the improvement in fit due to the lower–asymptote parameter that accounts for chance successes. The difference of two times the maximum likelihood from separate fittings of the model to the same data is distributed in large samples as a chi–square variable with degrees of freedom equal to the number of items. Applied to the first–stage data, these differences show mixed results as follows:

Subtest	*Chi-square*		
	Difference	df	Prob
Earth Sciences	28.4	6	<0.001
Biology	No change	6	1.000
Chemistry	97.4	6	<0.001
Physics	No change	6	1.000

Although we might have assumed different models for different subtests, for ease of presentation, and because rather small effects would be detected in the very large sample (N = 6675), we elected to assume the 2–parameter model throughout.

Results

Because the test forms at a given level of difficulty were assigned randomly, every examinee had the same probability of receiving any of the six stratified random forms. The test forms could therefore be scaled comparably by equivalent groups equating, which does not require linking items between forms. This is the preferred method of forms equating in large–scale assessment because it allows the maximum number of distinct items in the instrument as a whole, and thus the greatest generalizability of the results in aggregate–level reporting. Our analysis involves estimation of the slope and threshold parameters of all items in the study based on equivalent groups equating between the forms. Non–equivalent groups equating is of course required between second–stage levels.

The analysis of the first–stage items is essentially a conventional one–group analysis, except that it makes use of the second–stage assignments of the students to estimate the latent proficiency distribution at each level, and the latent distribution for the sample as a whole. These distributions are then used in the analysis of the combined first– and second–stage data as described for prototype 1 in the appendix. The analysis was very extensive, requiring simultaneous estimation of item parameters in the first–stage test and the three second–stage levels in each subject area of the six random parallel forms. Even in the abbreviated one–period test, this came to 538 distinct items in the multiple test by–group–by–form IRT analysis performed with the BILOG–MG program (in 45 minutes of execution time on an IBM 166 megahertz laptop computer).

Details of this analysis are too voluminous to present here, but will appear in the SSAS report (Bock, Doran, and Zimowski, 1998). The resulting parameter estimates were not uniformly favorable for efficient two–stage testing. When we were constructing the second–stage tests, the lack of item statistics relevant to Ohio twelfth graders forced us to assign items to the difficulty levels by subjective judgment. In numerous instances, the analysis results showed the judgments to be inaccurate, mostly in underestimating item difficulty. (In a study of expert judgment as an alternative to empirical estimation of item difficulties in test construction, R. L. Thorndike, 1982, reported similar results.) These misplaced items necessarily reduced the effectiveness of the second–stage test.

In addition, the above described procedure for assigning students to second–stage levels placed relatively few in the difficult level:

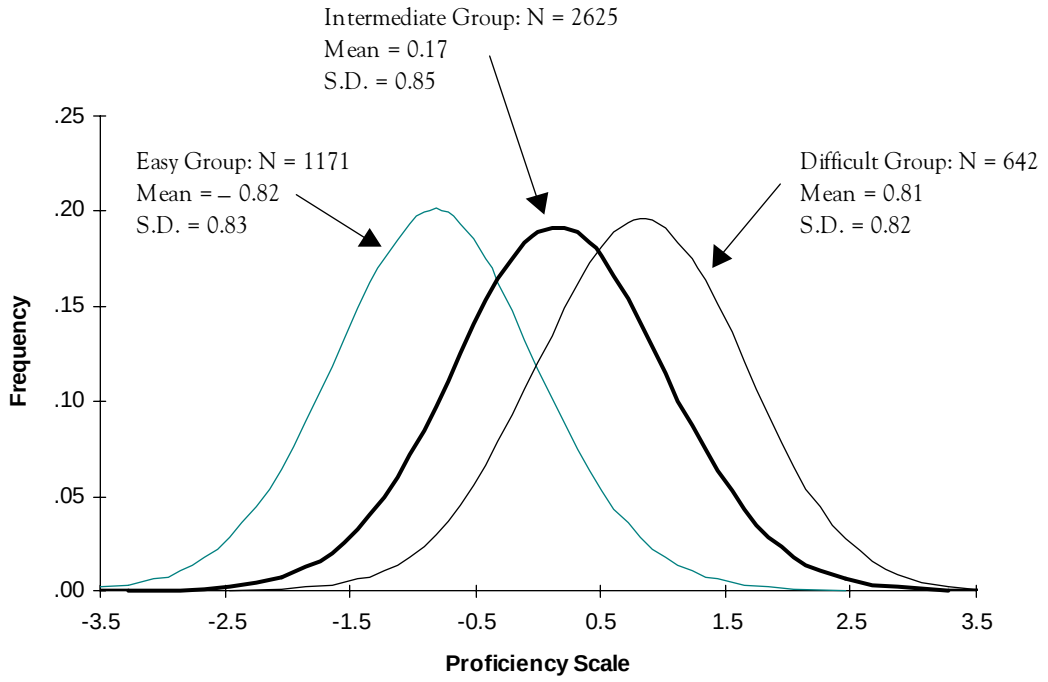
Area	*Level*		
	Easy	Intermediate	Difficult
Earth Sciences	1686	2294	459
Biology	1171	2625	642
Chemistry	2473	1581	398
Physics	2823	1277	355

Although a realistic appraisal of student preparation, these splits between levels were unfavorable to accurate estimation of item parameters in the difficult level, where sample sizes per item were very small.

Moreover, the small number of items in the first–stage test in each science area, even when augmented by student course background as described above, did not accurately assign examinees to second–stage levels. This is apparent in Figure 1 showing the estimated latent distributions among the groups of students assigned to the three levels of the Biology second–stage test. The distributions overlap considerably more than the corresponding distributions of spelling proficiency shown in Figure A–3 of the appendix, where the assignment was based on 16 first–stage items. The course background information was of little help: most students had taken the required first course in biology but nothing further. The distributions for chemistry shown in Figure 2, offer an interesting contrast, however. Since chemistry is an elective in Ohio high schools, and probably chosen only by students interested in a college–track science course, there is a strong separation between intermediate and easy levels, which reflects the presence or absence of one course. The first–level test items are also more discriminating than in biology because they require specialized knowledge that is available to most students only through a high school chemistry course. That is somewhat less true of biology, which figures more prominently in middle school science instruction.

These results suggest, however, that the number of pretest items in each science area should have been greater—eight or ten instead of six. More items would justify the use of IRT scale scores, rather than number–correct scores, for assignment of students to the second–stage booklets. This would have improved prediction for Earth Sciences and Biology where course background was an ineffective screen. Also, considering the importance of the first–stage items to the efficiency of the two–stage testing, we would like to have constructed the first–stage tests from among items pretested in the target population. Unfortunately, the necessary item statistics were not available in the Ohio study.

Figure 1. Biology: Estimated Latent Distributions



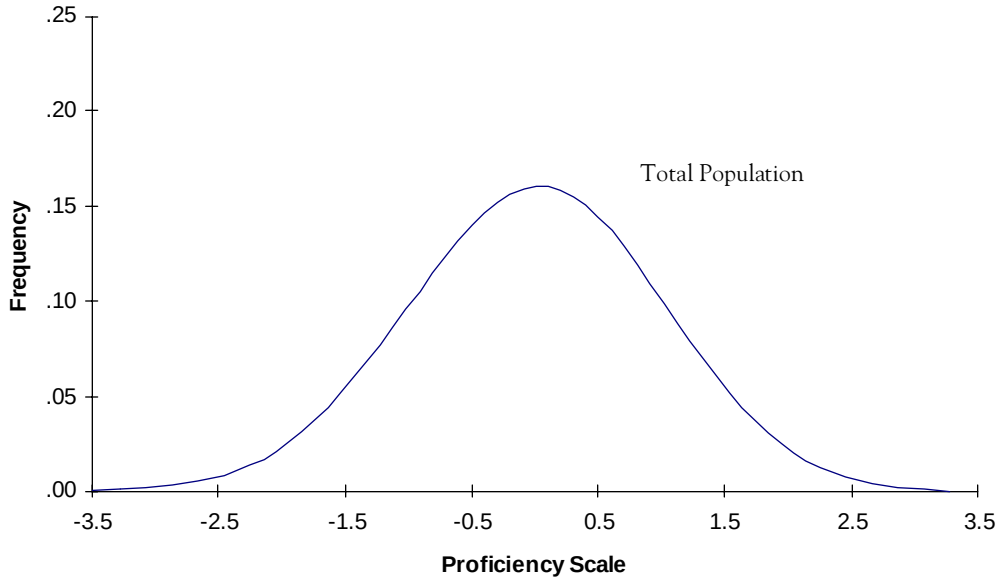
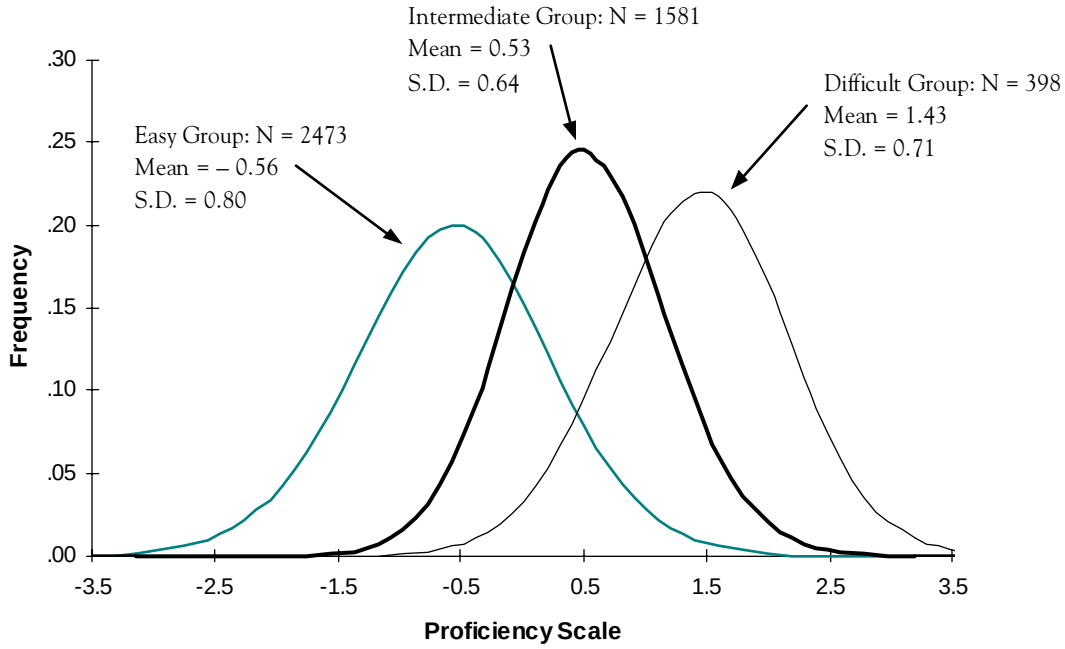
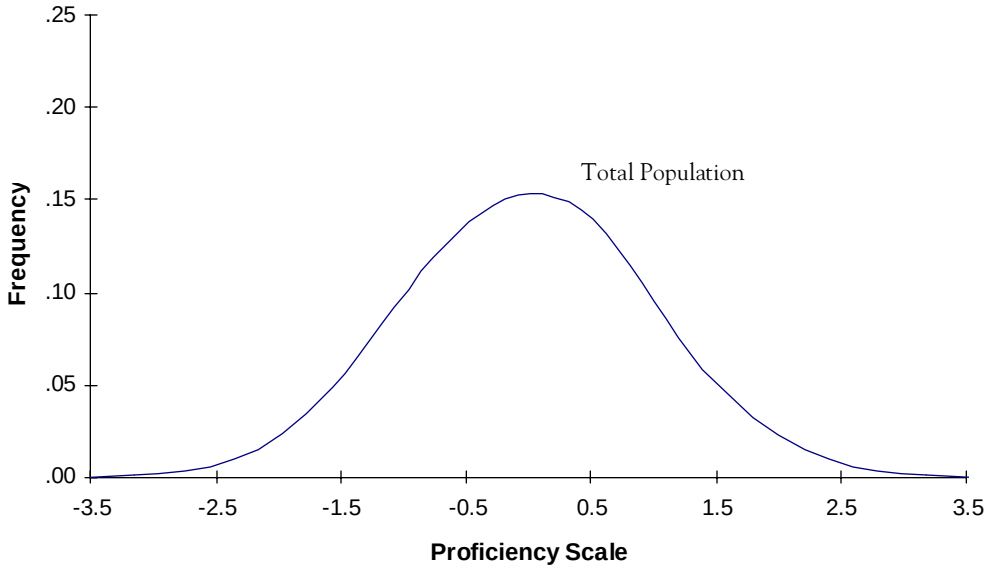


Figure 2. Chemistry: Estimated Latent Distributions





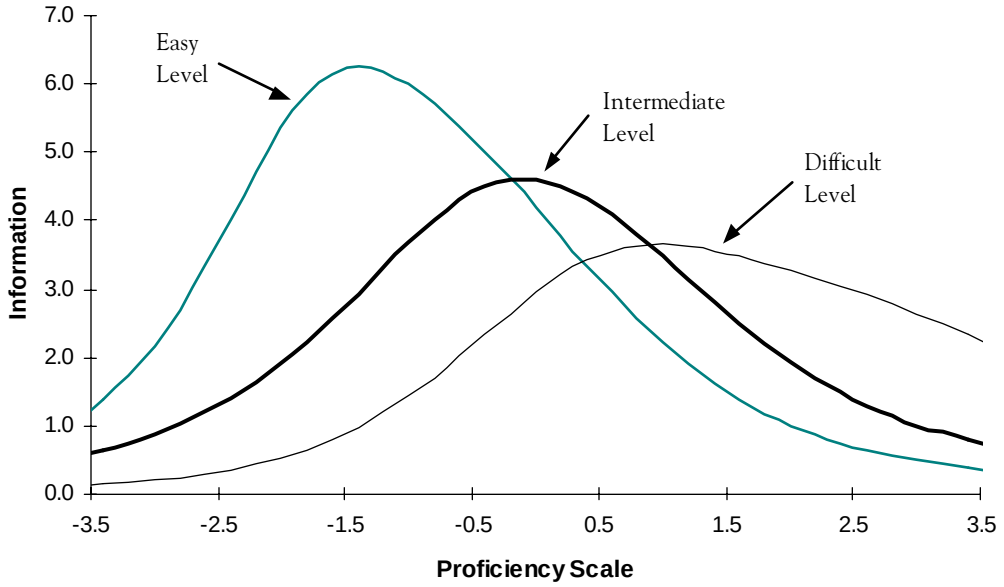
Information and efficiency. To assess the gains in efficiency attributable to the two–stage testing procedure, we computed relative information curves similar to those introduced by Birnbaum (1958). The formula for computing the values represented in these curves appears in the appendix of the present paper. The curves describe the precision (i.e., the reciprocal of the squared standard error of measurement) of maximum likelihood estimates of examinee proficiency as a function of proficiency level. Because a test–information curve is the sum of the constituent item–information curves, the precision and standard error of measurement for a test of any length can be predicted by multiplying the information of an actual test by the ratio of its length to that of a similar test of some other length. We make use of these predictions in extending the information values for the 6–item first–stage tests and the 8–item second–stage tests in the Ohio study to match the 16–item tests at both levels assumed for the two–stage prototypes in the appendix.

Since information and efficiency curves apply to test forms (i.e., test booklets) rather than tests, there will be separate curves for each science subject and each second–stage level of each test form. This amounts to 72 information curves (4 tests, each with 6 forms at 3 second–stage levels). For purposes of illustration we exhibit here only the curves for the first form of the biology and chemistry tests. These curves, which appear in Figure 3 through Figure 6, show the essential features of the other curves.

As discussed in the appendix, we would like to see information values between 5 and 10 in the regions of the proficiency scale spanned by the successive levels of the second–stage tests. On a scale in which the overall latent distribution has a standard deviation 1.0, as it does in Figure 2 and Figure 3, these values correspond to a test reliability of 0.8 and 0.9, respectively. Using this criterion, the biology test performs fairly well at the easy level, but less so at the intermediate level, and even less so at the difficult level. The problem is that the average discriminating power of the 134 biology items is only 0.566, considerably below averages in the neighborhood of 1.0 that are typical of achievement tests at the middle– and secondary–school grade levels. The chemistry test fares much better in this regard. The information curves in Figure 4 are above five for a wide range of the proficiency scale—from almost plus 1.5 to almost minus 1.5.

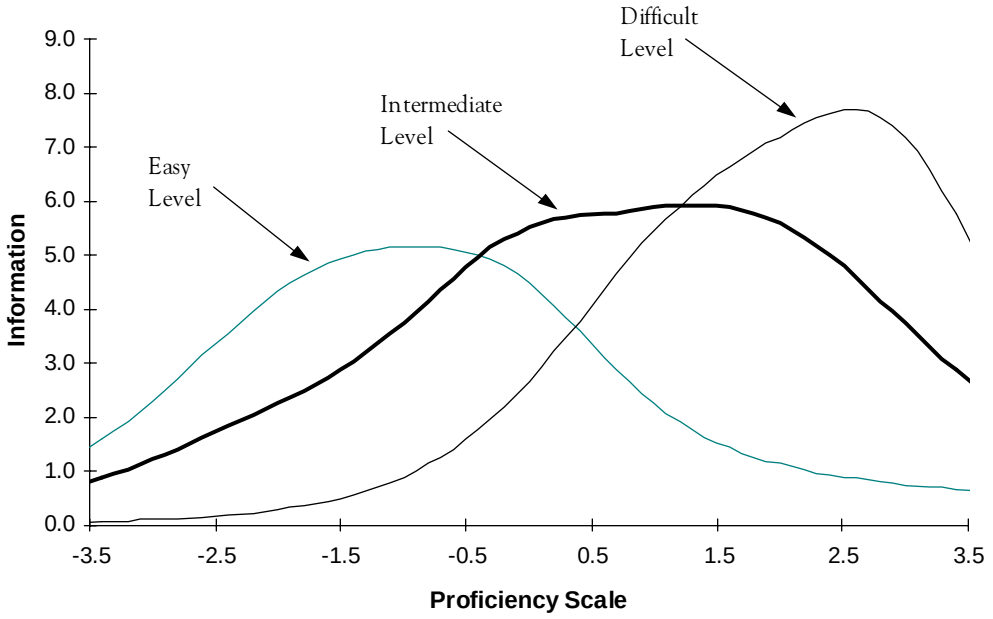
The respective efficiency curves for the biology and chemistry tests in Figure 5 and Figure 6 show a similar picture. Although we would like to see values greater than 2.0 everywhere, the biology test exceeds this criterion only at the extremes of the distribution where relatively few students would benefit from the increase in precision offered by the two–stage test. This is again a consequence of misplaced items in the second–stage tests, the misclassification of students by the first–stage measures, and the generally low discriminating power of the biology items. The results for the chemistry test are not any better, despite the stronger separation of the second–stage groups in this subject.

Figure 3. Biology: Two–stage Test Information



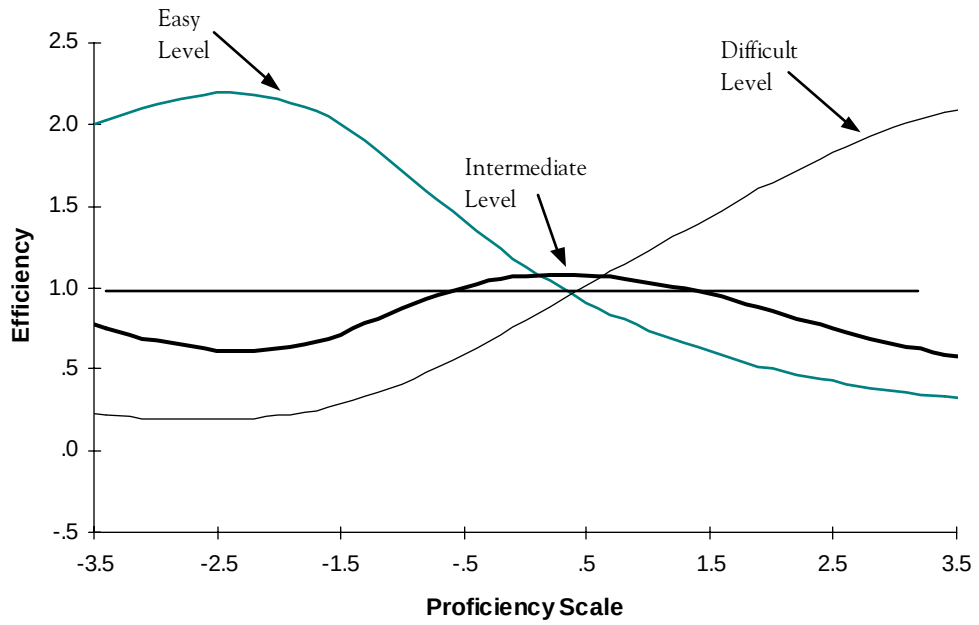
Assuming 16–item first– and second–stage tests.

Figure 4. Chemistry: Two–stage Test Information



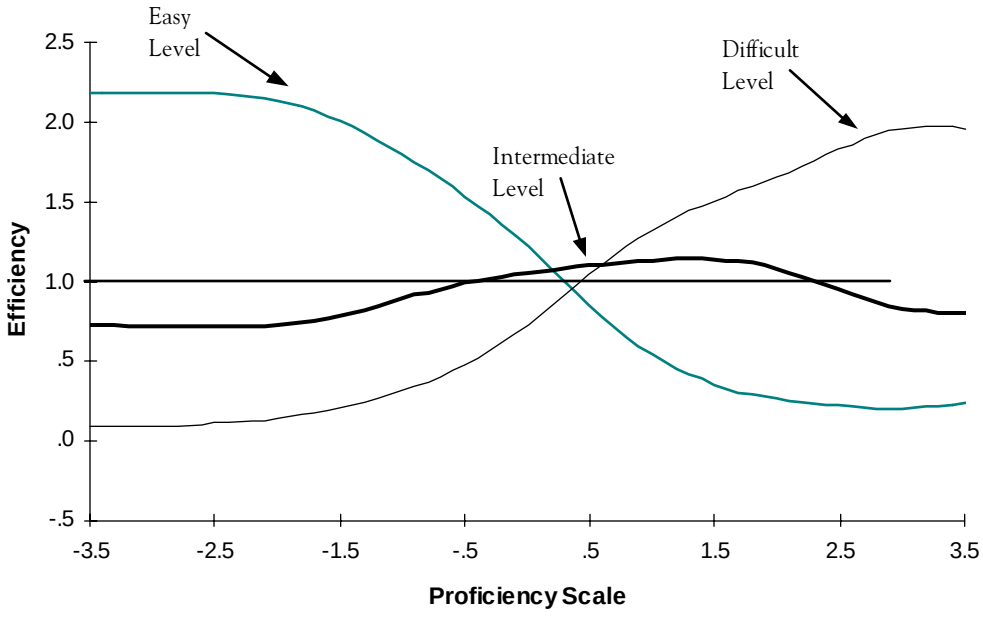
Assuming 16–item first– and second–stage tests.

Figure 5. Biology: Two–stage Test Efficiency



Assuming 16–item first– and second–stage tests.

Figure 6. Chemistry: Two–stage Test Efficiency



Assuming 16–item first– and second–stage tests.

Remarks on the Ohio study

The rather poor performance of the science tests that we constructed in the Ohio study can be attributed to the lack of item statistics obtained empirically in field trials in the population of students in question. This situation would not arise in an ongoing assessment program where variant items can be seeded into operational test forms to ensure a supply of pre–tested items. It does not detract, however, from the study as a demonstration that the field–procedures worked smoothly, the computerized production and assignment of personalized second–stage test forms was straightforward, and the analytical procedures for item parameter estimation and computation of science proficiency scores performed as expected. The study demonstrated that these techniques are in a sufficiently mature state to be applied in an operating assessment if suitable item banks are available.

The study also showed that two–stage testing can be carried out effectively by local school personnel when the test materials are well–organized and all data handling steps are performed centrally. In terms of the costs of two–stage versus one–stage testing, the marginal increase is limited to the production and handling of two test booklets per student, rather than one, but the total number of pages in these booklets will in general be smaller than in a one–stage test of equal precision.

Conclusions

The following conclusions answer the main questions in this study.

> *Is it possible, using modern data processing and document handling technology, to conduct a two–stage assessment in which first–stage booklets are returned to a central site where they are scored and then used to assign students to tailored, personalized second–stage booklets for a second testing session one or two weeks later?*

The field trial of a two–stage assessment of science attainment at school–leaving age, involving all twelfth–grade students in 40 Ohio schools, successfully used computer generated, laser printed copies of personalized tests to carry out two–stage testing in the four science areas. All test materials were shipped by overnight delivery from and to NORC in Chicago, where item analysis and test scoring were carried out. Although, as a first–time effort, the field trial required considerably more time between first– and second–stage testing, all operations were automated in a manner that would permit one–week turnaround of the first–stage documents in a working assessment.

> *Despite the confounding of item characteristics and respondent characteristics that is inherent in adaptive testing data, can valid IRT item calibration be carried out in a joint analysis of the nonadaptive first–stage data and the adaptive second–stage data?*

Simulation studies described in the appendix, based on actual test data, demonstrate that item calibrations of this type, which are essential in large–scale assessment programs, are possible. Further studies using larger samples of data should be undertaken to establish this result more firmly and to investigate the properties of this IRT parameter estimation in this context.

Can a two–stage assessment design attain the same generalizability of aggregate level scores as the NAEP matrix sampling design?

The study demonstrated that multiple randomly parallel test forms, stratified by the levels of the second–stage tests, can be calibrated on the same IRT scale by equivalent groups equating. This means that a level of aggregate generalizability equal to that of the present NAEP design will be attained if the total number of unique items per reporting area are the same in the two designs. Linking items are required between second–stage levels within forms but not between forms.

For purposes of reporting percentages of students at or above specified achievement levels, can the latent distribution of examinee proficiency be estimated accurately from two–stage data?

Comparison of latent distributions estimated in the simulated two–stage data, compared with one–stage estimates from the same data, confirmed the validity of the two–stage procedure. These latent distributions, estimated by multiple–groups maximum marginal likelihood estimation jointly with the item parameters, can model the latent distributions in NAEP demographic subgroups in ways equally or more accurate than the present plausible value method. These models include representing distribution densities by the Johnson and Kotz (1970) family of curves, as normal mixture distributions, or by kernel or spline smoothing over assigned points of support. To obtain the reporting percentages, the estimated densities may then be integrated up to the achievement level boundaries. This implies that plausible value imputation, which serves the same purpose in NAEP, could be replaced by computationally less intensive direct integration methods.

Can student–level scores sufficiently reliable for reporting for low–stakes uses by students, their teachers, and parents be obtained from relatively short two–stage tests?

Theoretical calculations reported in the appendix and confirmed in information analyses of the simulated two–stage data indicate that the combined data records of 16–item first– and second–stage tests will have reliabilities between 0.8 and 0.9 over a wide range of the score distribution. Reliabilities in this range are generally considered adequate for low–stakes purposes.

Compared to a one–stage test with the same measurement characteristics, does the two–stage test result in sufficient saving of testing time to justify the increased costs of test administration?

Although the item bases used in our simulation and in the Ohio study were not of a quality comparable to the NAEP item bases, we were able to demonstrate efficiencies of the two–stage testing in the neighborhood of 2.0, especially away from the mean of the population distribution. Closer to the mean, the information yield of the two–stage tests based on 16–item first– and second–stages is sufficiently high that high efficiency is not important. Because of the importance for policy purposes of information at both the high and low achievement levels, and considering the adverse effects of lengthy testing sessions on school recruitment and student cooperation, the reduction of testing time by one–half implied by these results could be justified as cost–effective. Even greater efficiencies should be possible with the item bases and item statistics available to NAEP.

Implications for NAEP and State NAEP

Implementation of adaptive testing procedures, and two–stage testing in particular, has the potential to increase the usability and validity of NAEP results by making available good quality student–level scores in main subject–matter areas. These scores would replace present plausible value scores as the basis for estimating percentages of students in the national and state populations who are at or above specified achievement levels. Because of their smaller measurement errors, the scores should provide more accurate estimates of the achievement level percentages, or equally accurate estimates with smaller sample sizes.

We do not suggest that adaptive testing in NAEP should be so extensive as to allow student level reporting of part scores within subject matters, such as the present six subscores in mathematics. Subscores would continue to be estimated at the aggregate level by plausible value methodology. However, the main subject–matter student–level score would be more effective than the student background characteristics for conditioning plausible value estimation, thus making the values more accurate and easier to compute.

Adaptive testing would help solve the two persisting problems for which NAEP has been most criticized—namely, lack of student motivation (see Bracy, 1997, for example) and failure to deliver assessment results in a timely manner. Adaptive testing would permit adequately reliable scores to be reported to individual students and their parents. With this personal stake in their performance on the NAEP tests, students should be less inclined to omit items, mark randomly, or give only token responses to writing or problem solving exercises. Overall gains in the NAEP score levels should result. In addition, by previously informing parents that scores will be reported to them, NAEP could increase community support for participation in the assessment and thus improve school–recruitment rates.

The improvement in data quality from adaptive testing would also speed data processing. The present procedure of conditioning on student background characteristics as a way of improving data quality would be replaced by the conditioning on first–stage test performance that is implicit and two–stage testing. Since the latter conditioning occurs during data acquisition rather than after all data have been collected and case weighted, a time consuming step in the data processing would be eliminated. Earlier reporting of the assessment results should then be possible.

An important side benefit of reporting is the student–level opportunity it presents for prospective studies of the validity of the assessment measures. Social survey agencies that specialize in longitudinal studies could work from the address database to locate samples of cases and interview them for their subsequent educational and occupational histories. The predictive value of the student–level scores in various subject matters could then be evaluated by standard statistical procedures. This information would be valuable as a contextual basis for objectively defining achievement–level standards. The practical implementations of the levels for access to and success in subsequent education or employment would give the achievement standards the consequential validity they now lack.

For state NAEP, the implications of adaptive testing very much depend on what role one assumes state NAEP should play in the educational policy decisions of state legislatures and departments of education. We suggest that the importance of state NAEP in this connection is not just the ordering of states by achievement levels, but

rather the possibility of evaluating the effects of major state–wide changes in curriculum, teacher qualifications, instructional materials, funding formulas, or local school governance and accountability, and the like. If the success of policy changes is to be judged by achievement outcomes, the state officials need the support of test results. This is particularly true in the case of policy changes that involve controversial issues such as bilingual education, whole language reading instruction versus phonics, or emphasis on calculator and computer use at the expense of basic computational skills. For credibility and dependability, these test results should come from a prestigious independent national source employing the most technically–sound sampling procedures and measurement methods. The results must also be regular, timely, and inclusive of main subject–matter areas at various grade levels.

State NAEP comes closer to filling these needs than any other national testing program. The needs cannot be met in testing conducted by programs using voluntary, nonrepresentative samples, or by special interest groups promoting particular views of education, or by private organizations whose testing methods and materials are not open to public inspection. Two–stage testing, or any other innovation in assessment, will therefore have implications for educational policy in the states to the extent that it improves the power of state NAEP to serve this impartial evaluative function.

In addition, there is another interesting application of state NAEP results that would result from the bench marks provided by good quality student–level scores. Many states have programs of achievement testing or assessment in which all public school students participate at selected grade levels. In these states, the students who have been selected for testing by state NAEP will also take the tests of the state's own program. If the student–level scores from state NAEP were made available to state departments of education in computer files also containing the student ID codes, that information could be merged with the student item response records on the state tests. The data would then be in a form suitable for a method of IRT test extension based on so–called "variant–item" technique (see Zimowski, Muraki, Mislevy and Bock, 1996). This procedure is the same as that used to estimate the parameters of new items introduced into operational test forms solely for field testing purposes. The variant items do not enter into the scoring of examinees, nor contribute to defining the construct that the actual items of the test specify. Instead, their parameters are estimated with respect to the construct and scale defined by the operational test items.

If this technique were applied to the combined item response records of the NAEP and state tests in the state NAEP subsample, the item discriminating powers of the state items estimated by extension from the NAEP national parameters would allow the information in the state items to be used to predict optimally the NAEP scores of all students taking the state tests. In this way, the state scores would be expressed on the national NAEP scale and achievement levels. In addition, the type of information analysis we have used here to evaluate the reliability and operating characteristics of two–stage tests could be applied to investigate the data quality of the resulting predicted scores.

Effectively, state NAEP would be providing behavioral bench marks, analogous to the physical bench marks of the U.S. Geological Survey, through which local test results could be linked to the NAEP–defined national norm. Similar proposals for linking state test results to NAEP have been around for some time, and committees and meetings have convened to discuss their feasibility and potential. With improved student–level data in state NAEP and modern IRT implementations of variant–item technique, the possibility of providing statistically rigorous national norms for state testing programs could become a reality.

References

Birnbaum, A. (1958). Further considerations of efficiency in tests of a mental ability. *Technical Report No. 17*. Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas.

Bock, R. D. (1997). The nominal categories model. In van der Linden & Hambleton, (Eds.). *Handbook of Item Response Theory*, pp. 33–49. New York: Springer–Verlag.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, 46, 443–445.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431–444.

Bock, R. D. & Zimowski, M. F. (1989). *Duplex design: Giving students a stake in educational assessment*. Chicago: National Opinion Research Center (NORC).

Bock, R. D. & Zimowski, M. F. (1997). Multiple group IRT. In van der Linden & Hambleton, (Eds.). *Handbook of Item Response Theory*, pp. 433–448. New York: Springer–Verlag.

Bock, R. D., Doran, R. & Zimowski, M. F. (1998). The School Science Assessment Study (in preparation).

Bracy, G. M. (1996). Altering motivation in testing. *Phi–Delta Kappan*, 78, 251–252.

Kiplinger, V. L. & Linn, R. L. (1996). Raising the stakes of test administration: the impact on student performance on the National Assessment of Educational Progress. *Educational Assessment*, 3, 111–133.

Linn, R. L., Rock, D. A. & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement* 29, 129–146.

Lord, F. M. (1971). A theoretical study of two–stage testing. *Psychometrika*, 36, 227–242.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores (with Contributions by A. Birnbaum)*. Reading, MA: Addison–Wesley.

O'Neill, H. F., Sugrue, B. & Baker, E. L. (1966). Effects of motivational intervention on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, 3, 135–157.

Owen, R. J. (1969). A Bayesian approach to tailored testing.Research Bulletin. No. 69–92. Princeton NJ: Educational Testing Service.

Rock, D. A. & Pollack, J. M. (1995). Psychometric report for NELS:88 base year through second follow–up. Report 95–382, Washington, DC: National Center for Education Statistics.

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*, pp. 309–317. New York: Academic Press.

Wainer, H. (Ed.) (1990). *Computerized adaptive testing: a primer*. Hillsdale, NJ: Earlbaum.

Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (1995). *BILOG–MG: Multiple–Group IRT Analysis and Test Maintenance for Binary Items.* Chicago: Scientific Software Int'l.

Appendix : Design and Analysis of a Two–Stage Test Instruments Suitable for NAEP and State NAEP

Educational assessment in NAEP differs from typical achievement testing programs in the extent and frequency of item updating in successive evaluations of the same subject matter. For this reason, the items in each proficiency scale are estimated in the current national data sample. This presents a problem for adaptive testing because, in data from typical adaptive testing sessions, item characteristics are confounded with the proficiency levels of the respondents. Indeed, in computerized adaptive testing, it is impossible to estimate parameters of items from data in most of the adaptive sequence because the observed proportions of correct response are too nearly uniform. For this reason, and also because of the expense of developing computerized tests, the item parameters are often estimated in paper–and–pencil versions on the assumption that their operating characteristics in the computer environment can be predicted from those values.

Precalibration of an item pool for adaptive testing does not seem practical for NAEP. The number of items involved, the large sample of examinees required, and the difficulty of duplicating the conditions of operational testing in the field trials militate against item parameter estimation in other than the current data. The studies we report here suggest, however, that multiple–group IRT estimation of item parameters in two–stage data is possible in the large sample sizes typical of state and national assessment. Depending on how strongly the first–stage test separates the latent distributions of proficiencies in the second–stage samples, the calibration procedure may have only a segment of the item response curve from which to estimate the parameters of items that appear in only one of the second–stage levels. However, with the strength added by priors on the guessing and slope parameters, the logistic response model should be conditioned well–enough in large samples to permit satisfactory estimation in two–stage data. We offer some evidence in support of this claim in an analysis of a simulated two–stage test based on data that allows comparison of two–stage and one–stage results. In this way, we evaluate two prototype two–stage analyses, the first of which we applied in the main text to data from the SSAS.

Prototype 1. A first–stage anchor–test design

For the first–stage test, the prototype described here assumes a block of 16 items devoted to a given subject matter. To allow rapid scoring, this test must be limited to multiple–choice items. A 45–minute administration time would therefore easily accommodate two subject–matter areas. Similarly, we assume 16–items in the second–stage block, some of which could be open–ended items requiring hand scoring. In light of Lord's (1971) finding of relatively small marginal gain between three and four levels, we assume three levels.

Assigning examinees to the second–stage levels will therefore require two cutting points on the proficiency scale. In order that the first–stage test discriminate maximally and equally well at these points, the 50–percent thresholds for correct response of eight of the items should be set at or near one of the cut–points, and the threshholds of the remaining items should be set at or near the other. As a rough guide to the discriminating power of a test peaked at two such points, let us assume that the slope parameters of the items (expressed in the metric of the normal ogive response model) are all equal to 1.0. Slopes in this range are relatively easy to obtain with

well–constructed multiple–choice items. We further assume that the number of alternatives in these items is 5, so that the probability of correct response by random marking is 0.2. On these assumptions we can compute the information function pertaining to the maximum likelihood estimate of proficiencies as follows:

Assuming the three–parameter logistic response model for item j, the probability of correct response is:

$$P_j(\theta) = c_j + (1 - c_j)\psi(z_j), \tag{1}$$

where

$$\psi(z_j) = \frac{1}{1 + \exp(-z_j)}$$

is the logistic function, and

$$z_j = Da_j(\theta - b_j)$$

is the corresponding logistic deviate or "logit." In the logit, $D = 1.7$ is the constant that converts the logistic metric to normal, c_j is the chance success parameters, a_j is the item discriminating power, θ is the examinee's proficiency value, and b_j is the location of the item on the proficiency continuum. As assumed above, we set $c_j = 0.2$ and $a_j = 1$ for purposes of discussion.

For the assumed three–parameter logistic model, the information for maximum likelihood estimation of θ conveyed by the response to item j is

$$I_j(\theta) = D^2a_j^2 \frac{1 - P_j(\theta)}{P_j(\theta)} \left(\frac{P_j(\theta) - c_j}{1 - c_j} \right)^2 \tag{2}$$

(see Lord, 1980, p. 73.)

The information about θ contained in the item response pattern of an n–item test is the sum of the item information:

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta). \tag{3}$$

The corresponding measurement standard error function for the test is the reciprocal square root of the information function:

$$SE(\theta) = I^{-1/2}(\theta) \tag{4}$$

It is easy to show that the q value corresponding to the maximum of information for item j is

$$\theta^* = b_j + \frac{1}{Da_j} \ln \frac{1 + (1 + 8c_j)^{1/2}}{2} \qquad (5)$$

which equals b_j + 1.57 on the above assumptions; that is, the point of maximum information is displaced upward somewhat by the chance success probability of multiple–choice items (see Lord, 1980, p. 152). After we have determined the location of the cutting points, we will choose items for the first–stage tests so that their points of maximum information cluster about those points.

A basis for choosing the points exists in the reasonable assumption that the population distribution of proficiency is normal. For convenience, we set the mean and standard deviation provisionally at 0 and 1, respectively. (In the NAEP assessments these values are set to 250 and 50 in the first assessment year.) On this provisional scale, we propose to locate the cutting points symmetrically at $\pm v$. Lacking any rationale for optimal locations, all other workers have set the points at equal percentiles of the distribution (i.e., at the 33.33 and 66.67 percentile points). This choice implies v = 0.426. In a large scale assessment context, however, we believe these points are too close to the mean. For the following reasons, we have set the points at the 25.00 and 75.00 percentiles, or v = 0.675. First, if item responses to the first– and second–stage tests are combined in estimating the examinee's proficiency, the information function of the middle–level test form is augmented by both components of the first–stage test; a larger proportion of the examinees therefore benefit from greater precision of that test. Second, a broader middle–level form permits us to move the point of maximum discrimination of the lower– and upper–level forms further toward the tails of the population distribution, which are of special interest in the policy uses of assessment. Third, it is important that very few students who belong in the lower level would be erroneously classified in the top level, and vice–versa—the wider middle–level proficiency interval assures very low probability of misclassification.

Given the assumed spacing, we can easily calculate the misclassification probabilities if we assume the errors of estimating q to be normally distributed about its true value and we know the standard error of measurement of the first–stage test at $+v$ and $-v$. For a rough estimate of the standard error, we first simplify (4) by neglecting the guessing effect, then compute the information for eight–item tests peaked at either cutting point;

$$I = 1.7^2 \times 8 \times 0.5 \times 0.5 = 5.78;$$

$$SE = (5.78)^{-1/2} = 0.416 .$$

Thus, there are $2 \times 0.675/0.416 = 3.25$ standard errors between the cutting points. The probability that an examinee with true proficiency in the neighborhood of the lower point would score above the upper point is therefore 0.0006, or odds of about one in a thousand. If equal percentiles are assumed as in Lord's (1970) study, the odds are 20 in a thousand. The smaller risk is preferable in sample sizes as large as NAEP's.

The next problem is where to place the maximum information points of the second stage tests. Obviously, if we assume peaked tests, the location of the maximum information of the middle–level test form should be located at zero on the proficiency scale. For good measurement in the tails of the population distribution, the maximum information of the lower– and upper–level forms should be well removed from zero. This requirement must be balanced, however, against the difficulty of precise estimation of item parameters when the success and failure rates for the items (*p*–values) are extreme. Having these forms peak beyond ±1.5 on the scale is unlikely to be productive in practical work. We have therefore adopted the values –1.5, 0.0, and +1.5 as the peaks of the second–stage forms.

Information curves for the first– and second–stage tests peaked in this way, and including the effect of the 0.2 rate of correct response attributable to random marking, are shown in Figure A–1. The curves for proficiencies estimated from responses to items in both stages are shown in Figure A–2. Notice the asymmetry of the curves and the influence of the first–stage test on their shape. Notice also that information yield falls off very sharply beyond about 2.0 and –2.25, but is still much improved over that of a test peaked at the mean. Better accuracy in the tails of the distribution is the main advantage of adaptive testing. In practice, the forms may not be this strongly peaked, and the curves will be broader and have lower maxima.

Figure A–1. Prototype 1: First- and second-stage information curves

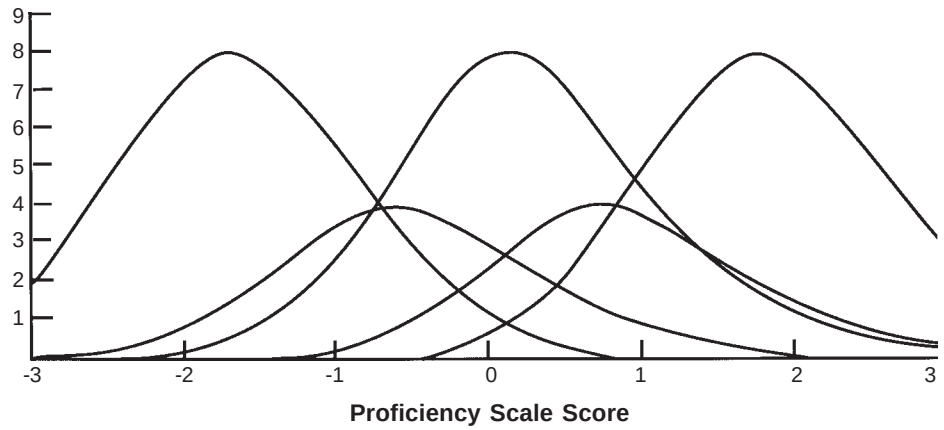
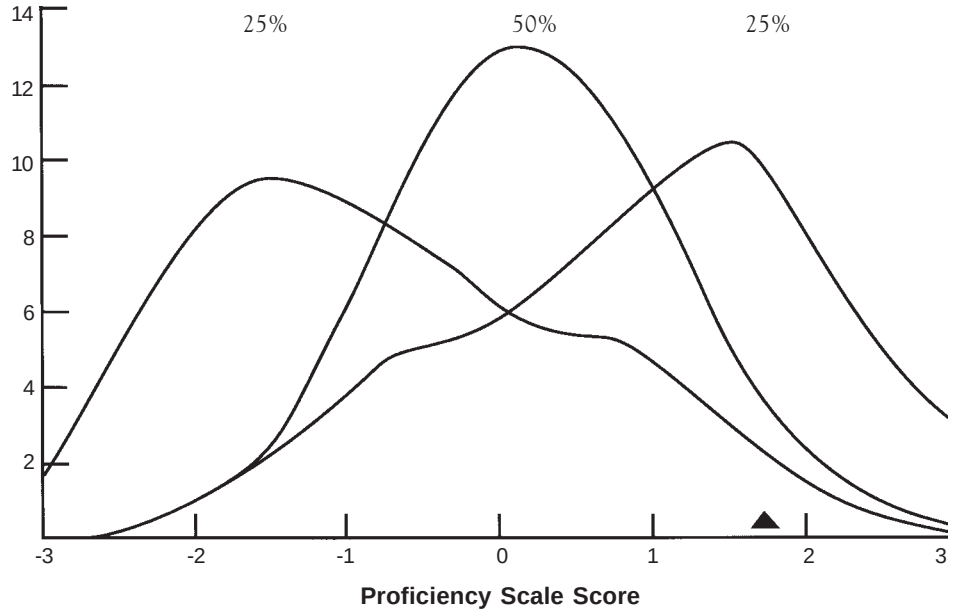


Figure A–2. Prototype 1: Combined first- and second-stage information



IRT analysis and scoring of prototype 1

To recover some of the information outside the range of the proficiency distributions in the second–stage levels, we propose a method of implementing multiple–group maximum marginal likelihood estimation that combines the first– and second–stage data. The steps are as follows:

1. First, estimate the respective latent distributions of proficiency among persons assigned to each of the second–stage levels. The distributions may be represented by weights at evenly spaced points on the proficiency continuum. Initially, the weights sum to unity within each distribution, but they are later combined proportionally to the corresponding sample sizes to form an estimate of the latent distribution of the whole population. To obtain these estimates, we need information from items that have been assigned nonadaptively to all respondents: the first–stage items serve this purpose. Because these items were previously calibrated in order to assign the respondents to second–stage levels, their parameters are available to the second–stage analysis.They can be used to estimate the posterior expected number of respondents in each level at specified points on the proficiency continuum assuming a standard normal prior distribution for the combined levels under the restriction that the mean and standard deviation of the combined distribution is 0 and 1. Alternatively, if persons have been assigned to the second–stage levels using other information in addition to the first–stage item responses, the latent distributions and first–stage item parameters may be estimated on the same assumptions after such assignment.
2. Using these estimated latent distributions as prior distributions, the parameters of items in the second–stage tests may be estimated, along with improved posterior estimates of the latent distributions, in a multiple–group maximum marginal likelihood IRT analysis. If the second–level test booklets do not include common linking items between levels, or if the linking is sparse, the first–stage items may be included in the analysis to provide anchor–test links.
3. Finally, given the estimated parameters for all first– and second–stage items, the proficiency scale scores for the respondents may be estimated by maximum likelihood or by Bayes using the corresponding latent distributions from the previous step as second–stage priors.

Example 1. A prototype–1 two–stage spelling test

As a small computing example, we simulated two–stage testing in data for the "One–Hundred Word Spelling Test" previously analyzed by Bock, Thissen, and Zimowski (1997). On the basis of item parameters they report, we selected 12 first–stage items and 12 items for each of three levels of the second–stage test. Because of the limited number of items in the pool, we could not meet exactly the requirements of the prototype design, but the resulting test illustrates well enough the main features of the analysis. The item numbers in this and a later example correspond to the words presented in Bock, Thissen, and Zimowski's Table 1. All computations in the analysis were carried out with the BILOG–MG program of Zimowski, Muraki, Mislevy and Bock (1996). The program command files for the analysis are included at the end of this appendix.

For assigning the cases in the data to second–stage levels under conditions that would apply in an operational assessment, we re–estimated the parameters for the 12 first–stage items, computed Bayes estimates of proficiency scale scores, and rescaled the scores to mean 0 and standard deviation 1 in the sample. Cases with scores at or below 0.67 were assigned to group 1, those at or above +0.67 were assigned to group 3, and the remainder to group 2. Of the 1000 cases in the original study, 274, 451, and 275 were assigned to groups 1, 2, and 3, respectively. With these assignment codes inserted in the case records, the latent distributions were estimated using the command file for the first–stage analysis shown below.

For the second–stage analysis, we used the latent distributions estimated in the first–stage analysis as the prior distributions for maximum marginal likelihood analysis of the combined first– and second–stage data. The points and weights representing the distributions are shown in the corresponding BILOG–MG command file. Inasmuch as there are no second–stage link items in this example, we use the first–stage items as an anchor test. The six easiest of these items provide the links between levels 1 and 2; the six most difficult provide the links between levels 2 and 3. The item parameter estimates resulting from this analysis are shown in Table A–1.

Because the spelling data contain responses of all cases to all items, we can examine the comparative accuracy of the estimates based on the 24 items per case in the two–stage data with those based on 48 items per case in a conventional one–stage test. The latter estimates are also shown in Table A–1. Despite the small number of items and relatively small sample size in this computing example, the agreement between the estimates is reasonably good for the majority of items. There are notable exceptions, however, among the second–stage items: of these, items 6, 7, 77, and 84 show discrepancies in both slope and threshold; all of these are from level 3 and have extremely high thresholds in the one–stage analysis, well beyond the +1.5 maximum we are assuming for second–stage items. Items 12 and 17 from level 3 are discrepant only in slope, as are items 26 and 38 from level 2, and items 50 and 64 from level 1. In all cases the two–stage slope is larger than the one–stage slope; this effect is balanced however, by the tendency of the first–stage items, 1, 4, 8, 10, 23, 25, 28, 29, 39, 47, 59, and 87 to show smaller slopes in the two–stage analysis. As a result, the average slope in the two–stage results is only slightly larger than the one–stage average. The average thresholds also show only a small difference. In principle, the parameters of a two–parameter logistic response function can be calculated from probabilities at any two distinct, finite values on the measurement continuum; similarly, those of the three–parameter model can be calculated from three such points. This suggests that in fallible data estimation must improve, even in the two–stage case, as sample size increases. Some preliminary simulations we have attempted suggest that with sample sizes in the order of 5 or 10 thousand, and better placing of the items, the discrepancies we see in the prototype 1 results largely disappear.

Table A–1. Prototype 1: Comparison of two–stage and one–stage item parameter estimates in spelling data (N = 1000)

	Two-Stage		*One-Stage*	
Item	**Slope S.E.***	**Threshold S.E.***	**Slope S.E.***	**Threshold S.E.***
1	0.686 0.072*	–0.179 0.071*	0.820 0.068*	–0.347 0.056*
4	0.704 0.071*	–0.514 0.068*	0.733 0.062*	–0.597 0.067*
5	0.779 0.134*	–1.488 0.113*	0.696 0.063*	–1.461 0.113*
6	0.769 0.132*	1.566 0.118*	0.291 0.039*	2.289 0.326*
7	0.702 0.147*	2.722 0.310*	0.332 0.054*	4.015 0.619*
8	0.516 0.062*	0.470 0.089*	0.517 0.049*	0.575 0.091*
9	0.794 0.126*	–0.230 0.079*	0.560 0.050*	–0.220 0.075*
10	1.031 0.085*	0.427 0.050*	0.957 0.070*	0.477 0.054*
12	0.473 0.103*	1.090 0.154*	0.784 0.073*	1.387 0.102*
14	0.799 0.141*	–2.020 0.167*	0.694 0.080*	–2.045 0.180*
15	0.440 0.098*	0.095 0.130*	0.407 0.044*	0.066 0.097*
17	0.733 0.128*	1.473 0.116*	0.480 0.049*	1.540 0.159*
20	0.483 0.105*	0.733 0.193*	0.306 0.039*	0.845 0.165*
23	0.497 0.061*	0.908 0.108*	0.522 0.053*	0.933 0.108*
24	0.546 0.113*	–1.931 0.207*	0.404 0.049*	–2.145 0.253*
25	0.644 0.067*	0.762 0.080*	0.703 0.060*	0.832 0.078*
26	0.531 0.111*	–1.688 0.177*	0.260 0.039*	–1.876 0.306*
27	0.733 0.121*	–0.105 0.080*	0.678 0.056*	–0.116 0.063*

Table A–1. Prototype 1: Comparison of two-stage and one-stage item parameter estimates in spelling data (cont)

	Two-Stage		*One-Stage*	
	Slope S.E.*	**Threshold S.E.***	**Slope S.E.***	**Threshold S.E.***
Item				
28	0.511 0.062*	0.436 0.090*	0.566 0.052*	0.501 0.082*
29	0.798 0.076*	–0.999 0.076*	0.840 0.078*	–1.029 0.081*
33	0.621 0.114*	–0.031 0.092*	0.556 0.051*	–0.098 0.074*
34	0.807 0.126*	–0.127 0.074*	0.740 0.062*	–0.044 0.059*
35	0.464 0.109*	2.649 0.370*	0.389 0.056*	2.856 0.373*
38	0.509 0.108*	–1.486 0.162*	0.373 0.045*	–1.639 0.207*
39	0.898 0.080*	–0.787 0.061*	0.846 0.071*	–0.880 0.069*
46	0.668 0.130*	–2.238 0.224*	0.747 0.090*	–2.135 0.185*
47	0.485 0.061*	0.489 0.095*	0.501 0.049*	0.555 0.093*
48	0.783 0.125*	–0.232 0.080*	0.636 0.054*	–0.096 0.067*
49	0.824 0.126*	0.043 0.071*	0.680 0.057*	0.035 0.063*
50	1.264 0.151*	0.038 0.049*	0.890 0.065*	–0.054 0.052*
53	0.487 0.106*	–1.256 0.153*	0.668 0.065*	–1.163 0.097*
54	0.684 0.119*	0.317 0.099*	0.810 0.068*	0.282 0.058*
59	0.614 0.067*	–0.649 0.080*	0.724 0.062*	–0.697 0.071*
60	0.936 0.146*	1.575 0.101*	0.609 0.060*	1.718 0.151*
64	0.415 0.095*	–0.222 0.143*	0.267 0.038*	–0.188 0.145*

Table A–1. Prototype 1: Comparison of two-stage and one-stage item parameter estimates in spelling data (cont)

	Two-Stage		*One-Stage*	
	Slope S.E.*	**Threshold S.E.***	**Slope S.E.***	**Threshold S.E.***
Item				
68	0.723 0.132*	–1.977 0.173*	0.515 0.063*	–2.079 0.220*
69	0.648 0.121*	1.510 0.132*	0.550 0.054*	1.540 0.141*
72	0.674 0.119*	–0.348 0.101*	0.592 0.051*	–0.295 0.072*
73	0.555 0.112*	1.565 0.155*	0.252 0.039*	1.947 0.322*
77	0.904 0.147*	1.844 0.128*	0.265 0.041*	3.073 0.476*
78	0.636 0.122*	–1.830 0.171*	0.623 0.069*	–1.835 0.161*
84	1.191 0.221*	2.566 0.203*	0.581 0.089*	3.493 0.436*
85	0.684 0.127*	–1.036 0.110*	0.419 0.046*	–1.114 0.145*
86	0.407 0.097*	2.149 0.291*	0.218 0.037*	2.633 0.465*
87	0.473 0.061*	–0.763 0.105*	0.518 0.052*	–0.800 0.101*
90	0.846 0.143*	–1.846 0.139*	0.792 0.084*	–1.827 0.139*
95	0.754 0.131*	–1.338 0.107*	0.600 0.059*	–1.280 0.118*
97	0.422 0.109*	3.431 0.593*	0.211 0.043*	5.217 1.053*
Average	0.678	0.074	0.565	0.224

The latent distributions estimated with items from both stages are depicted in Figure A–3. The distributions for the three assignment groups are shown normalized to unity. The estimated population distribution, which is the sum of the distributions for the individual groups weighted proportional to sample size, is constrained to mean 0 and standard deviation 1 during estimation of the component distribution. It is essentially normal and almost identical to the population distribution estimated in the one–stage analysis.

One may infer the measurement properties of the simulated two–stage spelling test from the information and efficiency calculations shown in Figure A–4 and Figure A–5, respectively. When interpreting information curves, the following rules of thumb are helpful. An information value of 5 corresponds to a measurement error variance of 1/5 = 0.2. In a population in which the score variance is set to unity, the reliability of a score with this error variance is 1.0 – 0.2 = 0.8. Similarly, the reliability corresponding to an information value of 10 is 0.9. In the context of low–stakes score reporting, we are aiming for reliabilities anywhere between these figures. As is apparent in Figure A–4, this range of reliability is achieved in the two–stage results for spelling over much of the latent distribution.

Finally, the efficiency curves in Figure A–5 for the three levels show us the saving of test length and administration time, including both first– and second–stage testing, due specifically to the two–stage procedure in comparison with a one–stage test of the same length and item content. In this case we hope to see efficiencies greater than 2.0, at least away from the population mean where conventional tests with peaked centers typically have reduced precision. The prototype 1 design and analysis meets this criterion.

To increase generalizability of group–level mean scores in assessment applications of the prototype 1 design, the second–stage tests will of course have to exist in multiple stratified randomly–parallel forms. As with matrix sampling designs, these forms will be administered in random rotation to the examinees in each second–stage level. The sample data will then be suitable for equivalent–groups equating of the second–stage forms.

Figure A–3. Prototype 1: Estimated latent distributions from two-stage and one-stage spelling data

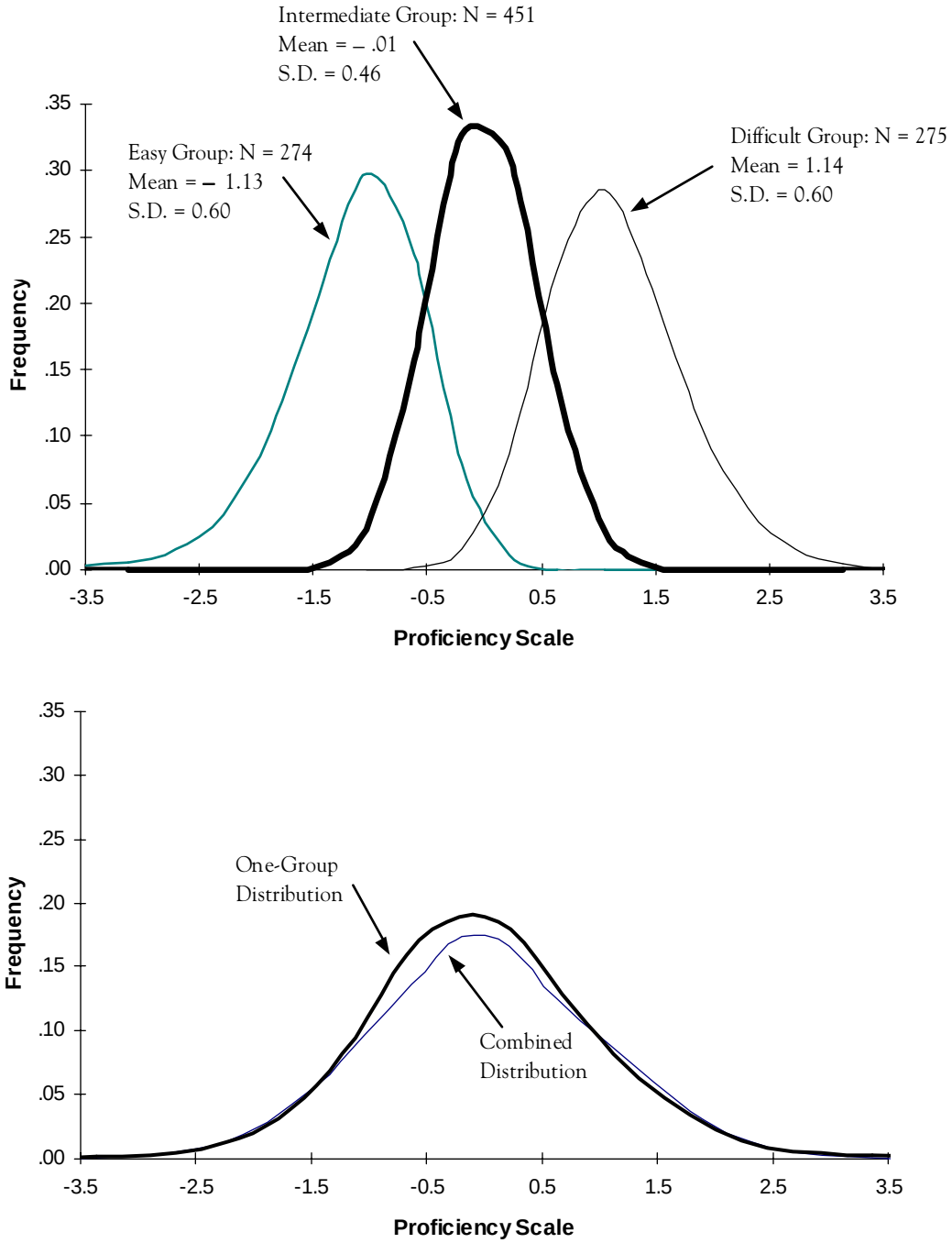
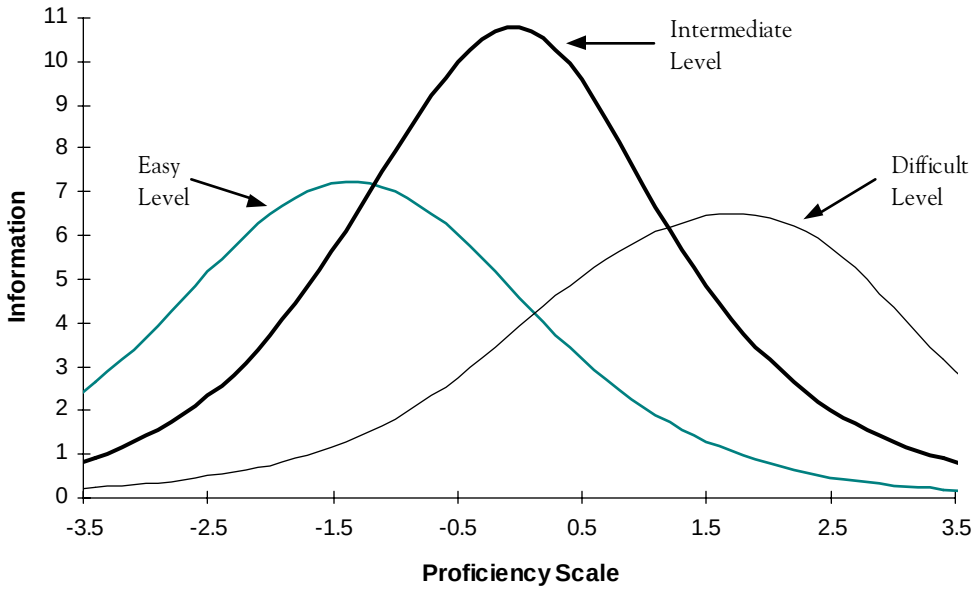
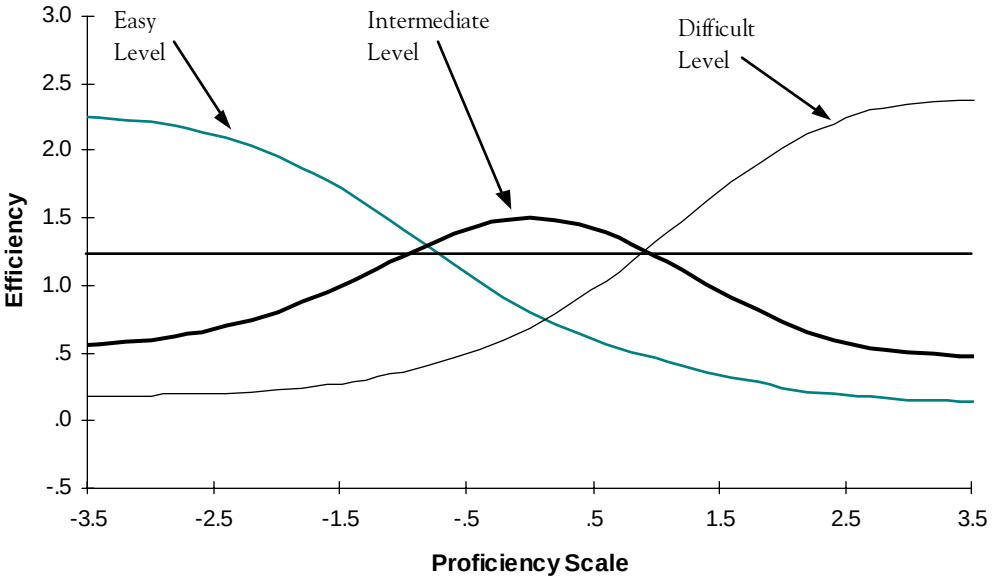


Figure A–4. Prototype 1: Two-stage spelling test information



Assuming 16–item first– and second–stage tests, plus information from 8, 16, and 8 first–stage items contributing to the Easy, Intermediate, and Difficult tests, respectively.

Figure A–5. Prototype 1: Efficiencies of the two-stage spelling tests



Assuming 16–item first– and second–stage tests, plus information from 8, 16, and 8 first–stage items contributing to the Easy, Intermediate, and Difficult tests, respectively; efficiencies relative to a one–stage test of the same length.

Prototype 2: An incomplete–block two–stage design

The prototype 2 design is a generalization of prototype 1 in which the first–stage test as well as the second stage test exists in multiple randomly–parallel forms. It assumes random assignment of the first–stage forms to examinees generally, and random assignment of second–stage forms within assigned second–stage levels. Although the type 2 design is not as efficient as the type 1 design, it has certain advantages which we will point out.

To re–deploy one of the second–stage forms as a type 2 instrument, we assume that there are three levels of difficulty and that the number of items at each level is divisible by 3. The $3n$ items may then be partitioned into blocks of size n, and the 3n–item second–stage tests represented as follows:

Levels	***Blocks***
1. Easy	a b c
2. Intermediate	d e f
3. Difficult	g h i

Now form three first–stage test booklets from the second–stage item blocks as follows:

Test	
A	a d g
B	b e h
C	c f i

Next construct three distinct 2–block second–stage test booklets within each level as follows:

Levels	***Blocks***
1. Easy	bc ac ab
2. Intermediate	ef df de
3. Difficult	hi gi gh

Any given examinee is assigned one of the following nine combinations of blocks in the first– and second–stage booklets. The blocks in the second are the "aliases" of those in the first, so that no examinee is presented the same item twice.

First-stage	Second-stage	1st-stage	2nd-stage
Test	**Level**	**Blocks**	**Blocks**
A	1	adg	bc
B	1	beh	ac
C	1	cfi	ab
A	2	adg	ef
B	2	beh	df
C	2	cfi	de
A	3	adg	hi
B	3	beh	gi
C	3	cfi	gh

In an actual assessment we would propose six items per block for each subject matter—thus, 18 items per first–stage booklet and 12 items per second–stage booklet. In the spelling data, however, there were only enough suitable items for 4–item blocks. We assigned the following items to the nine blocks:

a. 5, 26, 68, 90
b. 14, 38, 78, 92
c. 24, 53, 85, 95
d. 9, 27, 48, 54
e. 1, 15, 33, 49
f. 4, 10, 34, 50
g. 6, 17, 25, 69
h. 8, 28, 35, 73
i. 12, 23, 47, 60

With more items in the first–stage than the second, this design cannot be as efficient as prototype 1, although it has the merit of allowing more time for extended–response items in the second stage. However, a more important advantage is that every item appears in one of the first–stage booklets. This means that all of the items can be calibrated in data from the nonadaptively administered first–stage testing. None of the problems associated with item calibration in adaptive testing will arise. If there is sufficient time between the two testings, the items for the second–stage forms can be selected and the booklets produced using the item parameters from the calibration in the first–stage data. This strategy would have been advantageous in the SSAS, where items were assigned to the second–stage forms with generally poor knowledge of their difficulties in the Ohio student population; unfortunately, we did not conceive of this approach at the time.

In an ongoing assessment program, the type 2 design would useful as a first step in in the direction of more efficient two–stage or computerized adaptive testing. Although data quality in the initial assessment would not be quite as high as in following assessments, the assessment program could get under way with less extensive pretesting than is required for conventional adaptive test construction. Once the

program was underway, all subsequent updating on the assessment instrument could be carried out in the operational data by inserting "variant" items in the first–stage forms as described in the body of the paper. In this way any further need for separate item field trials would be eliminated.

The setup of the BILOG–MG analysis for item parameter estimation in the prototype 2 first–stage data appears at the end of the appendix following the command files for the prototype 1. Notice that there are no common items between the three test forms. It is a common misapprehension that linking items are required in IRT equating of test forms, whereas linking is actually necessary only in non–equivalent groups equating. The prototype 2 simulation, we assigned the forms in rotation to the 1000 respondents, generating three groups sampled from the same population. The analysis differs from a one–group analysis only in that a given examinee responds to a random sample of 12 items, compared to the full set of 36 items in the one–group analysis; consequently the number of respondents per item is one–third that of the one–group analysis. The concordance of the two analyses is apparent in Table A–2: none of the differences between corresponding parameter estimates is excessive compared to their standard errors, and the mean slope and threshold are essentially the same. Similarly, the average latent distribution from the two–stage analysis, shown in Figure A–6, is essentially the same as that from the one–group analysis and to that of prototype 1 (Figure A–3).

Table A–2. Prototype 2: Comparison of two-stage and one-stage item parameter estimates in spelling data (N = 1000)

	Two-Stage		*One-Stage*	
	Slope S.E.*	**Threshold S.E.***	**Slope S.E.***	**Threshold S.E.***
Item				
1	0.827 0.168*	–0.207 0.097*	0.878 0.076*	–0.336 0.053*
4	0.819 0.144*	–0.662 0.119*	0.806 0.071*	–0.562 0.062*
5	0.999 0.212*	–1.328 0.182*	0.756 0.073*	–1.367 0.108*
6	0.301 0.091*	2.490 0.740*	0.301 0.041*	2.219 0.319*
8	0.525 0.111*	0.535 0.168*	0.525 0.052*	0.562 0.091*
9	0.555 0.113*	–0.336 0.138*	0.595 0.054*	–0.212 0.071*
10	0.903 0.150*	0.429 0.099*	0.989 0.075*	0.453 0.052*
12	0.662 0.150*	1.616 0.289*	0.800 0.075*	1.355 0.101*
14	0.570 0.199*	–2.407 0.655*	0.750 0.088*	–1.915 0.169*
15	0.343 0.095*	0.092 0.199*	0.409 0.047*	0.065 0.096*
17	0.529 0.199*	1.529 0.325*	0.493 0.051*	1.499 0.156*
23	0.623 0.120*	0.826 0.174*	0.539 0.055*	0.901 0.107*
24	0.365 0.103*	–2.402 0.640*	0.436 0.054*	–2.003 0.239*
25	0.811 0.157*	0.773 0.144*	0.712 0.062*	0.812 0.078*
26	0.201 0.092*	–2.263 1.042*	0.266 0.041*	–1.830 0.304*
27	0.566 0.122*	–0.079 0.128*	0.690 0.061*	–0.188 0.062*

Table A–2. Prototype 2: Comparison of two-stage and one-stage item parameter estimates in spelling data (cont)

	Two-Stage		*One-Stage*	
	Slope S.E.*	**Threshold S.E.***	**Slope S.E.***	**Threshold S.E.***
Item				
28	0.433 0.104*	0.570 0.200*	0.571 0.055*	0.491 0.082
33	0.597 0.125*	0.074 0.126*	0.608 0.056*	–0.097 0.068*
34	0.732 0.135*	–0.104 0.104*	0.757 0.064*	–0.048 0.058*
33	0.597 0.125*	0.074 0.126*	0.608 0.056*	–0.097 0.068*
34	0.732 0.135*	–0.104 0.104*	0.757 0.064*	–0.048 0.058*
35	0.535 0.134*	1.977 0.418*	0.401 0.057*	2.781 0.359*
38	0.637 0.136*	–1.049 0.201*	0.385 0.049*	–1.586 0.205*
47	0.590 0.115*	0.778 0.172*	0.525 0.052*	0.529 0.089*
48	0.831 0.157*	–0.122 0.096*	0.649 0.057*	–0.098 0.065*
49	0.711 0.140*	0.085 0.110*	0.711 0.061*	0.028 0.061*
50	1.269 0.255*	–0.030 0.073*	0.957 0.072*	–0.062 0.049*
53	0.705 0.162*	–1.138 0.198*	0.725 0.072*	–1.092 0.091*
54	0.746 0.160*	0.294 0.177*	0.854 0.073*	0.262 0.056*
60	0.504 0.144*	1.754 0.366*	0.591 0.060*	1.744 0.160*
68	0.393 0.129*	–2.316 0.677*	0.544 0.070*	–1.951 0.210*

Table A–2. Prototype 2: Comparison of two-stage and one-stage item parameter estimates in spelling data (cont)

	Two-Stage		*One-Stage*	
	Slope S.E.*	**Threshold S.E.***	**Slope S.E.***	**Threshold S.E.***
Item				
69	0.609 0.122*	1.413 0.268*	0.550 0.055*	1.528 0.146*
73	0.329 0.099*	1.243 0.399*	0.252 0.040*	1.948 0.311*
78	0.695 0.183*	–1.614 0.309*	0.672 0.073*	–1.720 0.148*
85	0.373 0.100*	–1.656 0.432*	0.426 0.048*	–1.093 0.145*
90	0.621 0.167*	–2.016 0.416*	0.808 0.093*	–1.772 0.149*
92	0.405 0.113*	–1.097 0.301*	0.527 0.056*	–0.871 0.104*
95	0.600 0.123*	–1.268 0.234*	0.638 0.066*	–1.214 0.133*
Average	0.609	–0.156	0.614	–0.077

Along with parameter estimation in the first–stage results, we computed case–by–case EAP scale–score estimates standardized to mean 0 and standard deviation 1 in the sample. Each case was then assigned to one of the three second–stage groups relative to the cutting points –0.67 and +0.67. The estimation of proficiency scores using combined information from the two stages was then carried out with the second prototype 2 BILOG–MG setup shown below. For convenience we used maximum likelihood score estimation, which does not require the input of weights from the first–stage latent distributions as does the EAP score estimation used in the prototype 1 scoring.

Average information curves for the 9 second–stage forms are shown in Figure A–7, extended from 20 to 32 items for purposes of comparison with the prototype 1 results. The information levels are high in the center of the latent distribution, but, as expected, fall off toward the tails more quickly than those of prototype 1. The problem is partly that the the first–stage item locations peak in the center rather than at the cutting points and there are too few easy and difficult items in the second–stage test. These conditions are unavoidable in the type 2 design. Similarly, the corresponding efficiency curves shown in Figure A–8 are less favorable in the tails. When evaluating the utility of the design, however, one must balance these results against the design's special merit in situations where accurate information on item difficulties is not available prior to the first–stage testing.

Figure A–6. Prototype 2: Estimated latent distributions from two-stage and one-stage spelling data

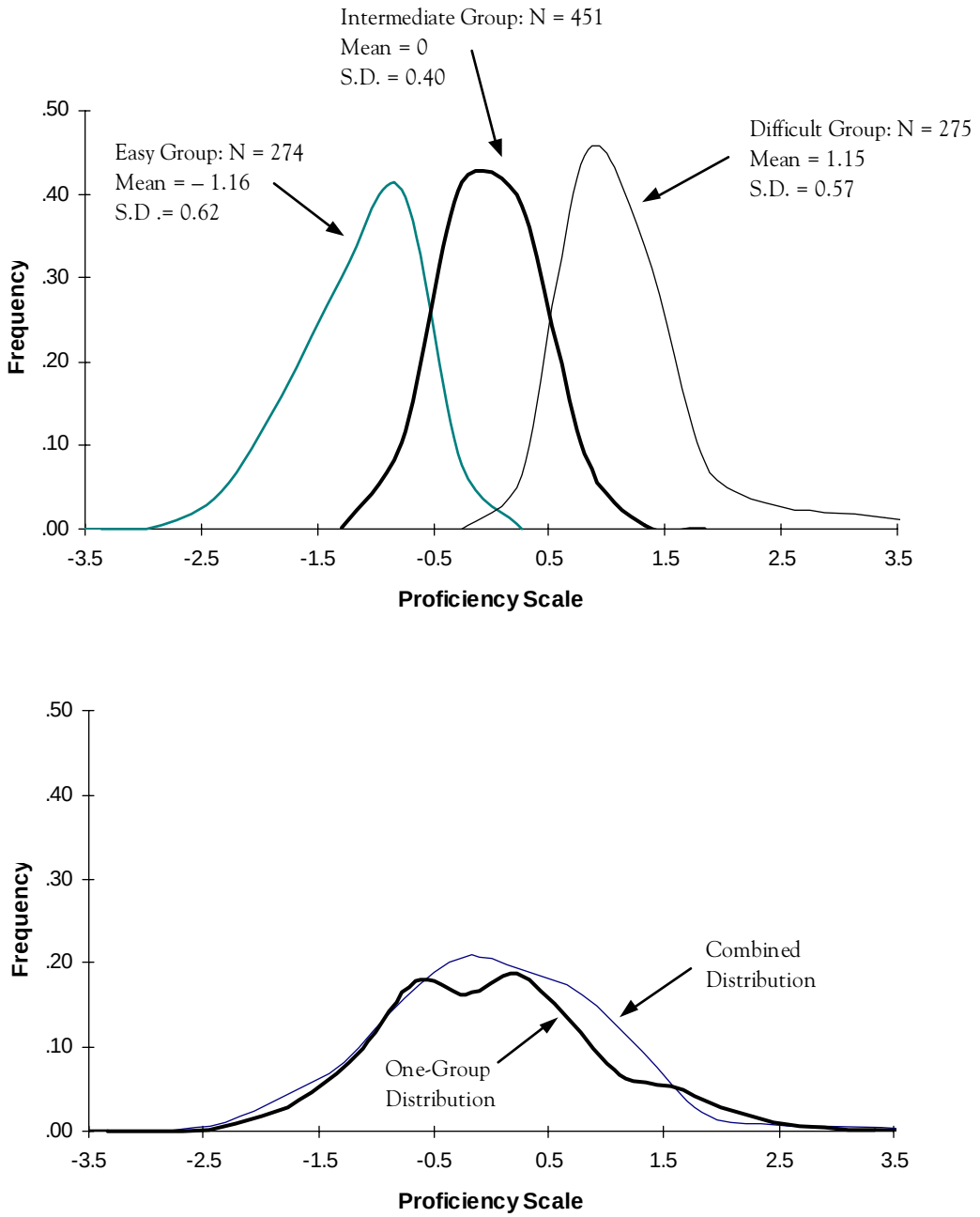
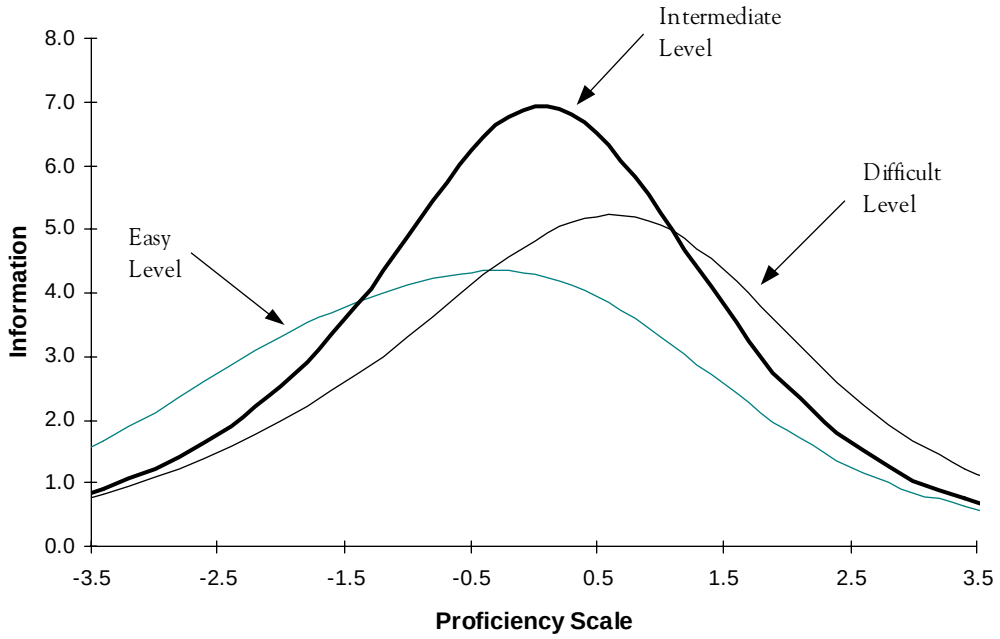
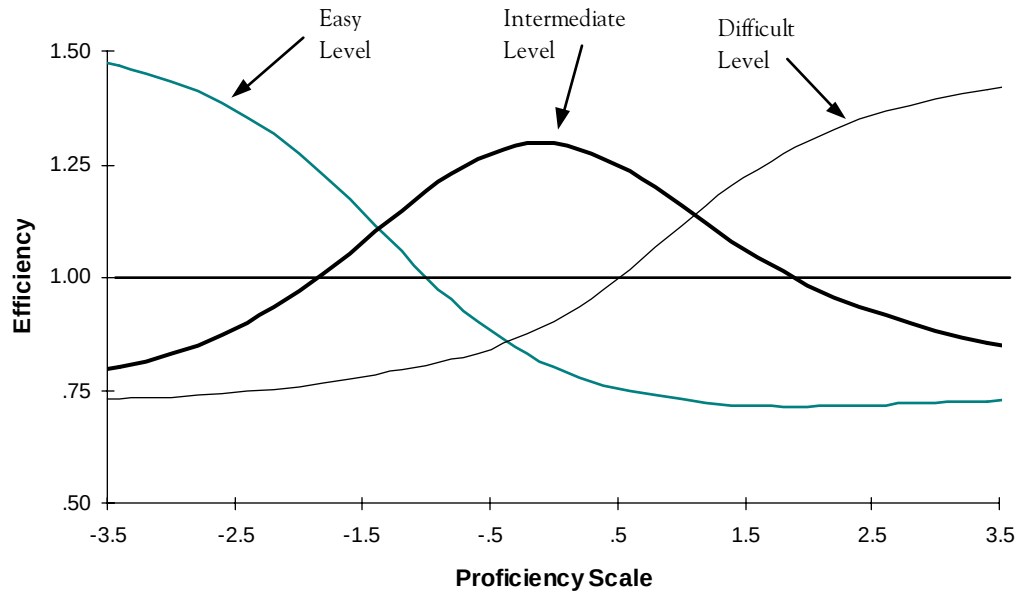


Figure A–7. Prototype 2: Two-stage spelling test information



Assuming a 15–item first–stage test and a 10–item second–stage test.

Figure A–8. Prototype 2: Efficiencies of the two-stage spelling tests



Assuming a 15–item first–stage test and a 10–item second–stage test.

BILOG–MG command files for the Prototype 1 computing example

```
ANALYSIS 1.: A SIMULATED TWO-STAGE SPELLING TEST
     Estimation of first-stage item parameters and latent distributions

>COMMENT
      Based on the 100-word spelling test data. N = 1000
      (See Bock, Thissen and Zimowski, 1997).

>GLOBAL   DFNAME='SPELLGRP.DAT',NPARM=2, SAVE;
>SAVE     SCORE='SPEL1ST.SCO',PARM='SPELL1ST.PAR';
>LENGTH   NITEMS=12;
>INPUT    NTOT=100, SAMPLE=1000, NGROUP=3, KFNAME='SPELLGRP.DAT',
          NIDCH=11,TYPE=1;

>ITEMS    INUM=(1(1)100), INAME=(SPELL001(1)SPELL100);
>TEST     TNAME=SPELLING INUM=(1,4,8,10,23,25,28,29,39,47,59,87);
>GROUP1   GNAME=GROUP1, LENGTH=12,INUM=(1,4,8,10,23,25,28,29,39,47,59,87);
>GROUP2   GNAME=GROUP2,LENGTH=12,INUM=(1,4,8,10,23,25,28,29,39,47,59,87);
>GROUP3   GNAME=GROUP3,LENGTH=12,INUM=(1,4,8,10,23,25,28,29,39,47,59,87);
          (11A1,I1,25A1,1X,25A1,/T13,25A1,1X,25A1)
>CALIB    FIX,NOFLOAT,CYCLE=35,SPRIOR,NEWTON=2,
          CRIT=0.001,REF=0;

>SCORE    IDIST=3,METHOD=2,NOPRINT,INFO=1,POP;

ANALYSIS 2: A SIMULATED TWO-STAGE SPELLING TEST
  Estimated link and second-stage item parameters, and latent
  distributions

>COMMENT
   Based on the 100-word spelling test data. N = 1000
   (See Bock, Thissen and Zimowski, 1997).

>GLOBAL   DFNAME='SPELLGRP.DAT',NPARM=2, SAVE;
>SAVE     SCORE='SPEL2N2.SCO',PARM='SPEL2N2.PAR';
>LENGTH   NITEMS=48;
>INPUT   NTOT=100,SAMPLE=1000,NGROUP=3,KFNAME='SPELLGRP.DAT',NIDCH=11,
          TYPE=1;
>ITEMS    INUM=(1(1)100), INAME=(SPELL001(1)SPELL100);
>TEST     TNAME=SPELLING,INUM=(1,4,5,6,7,8,9,10,12,14,15,17,20,23,24,25,
26,27,28,29,33,34,35,38,39,46,47,48,49,50,53,54,59,60,64,68,69,72,73,
77,78,84,85,86,87,90,95,97);
>GROUP1  GNAME=GROUP1, LENGTH=18,
INUM=(1,4,5,14,24,26,29,38,39,46,53,59,68,78,85,87,90,95);
>GROUP2  GNAME=GROUP2,LENGTH=24,INUM=(1,4,8,9,10,15,20,23,25,27,28,29,
33,34,39,47,48,49,50,54,59,64,72,87);
>GROUP3  GNAME=GROUP3, LENGTH=18,
INUM=(6,7,8,10,12,17,23,25,28,35,47,60,69,73,77,84,86,97);
 (11A1,I1,25A1,1X,25A1,/T13,25A1,1X,25A1)

>CALIB    IDIST=1,FIX,NOFLOAT,CYCLE=35,SPRIOR,NEWTON=2,CRIT=0.001
          REF=0,PLOT=1.0,ACC=0.0;
>QUAD1 POINT=(-0.4064E+01 -0.3636E+01 -0.3209E+01 -0.2781E+01 -0.2353E+01
             -0.1925E+01 -0.1497E+01 -0.1069E+01 -0.6415E+00 -0.2137E+00
              0.2142E+00  0.6420E+00  0.1070E+01  0.1498E+01  0.1926E+01
              0.2353E+01  0.2781E+01  0.3209E+01  0.3637E+01  0.4065E+01),
```

BILOG-MG command files for the Prototype 1 computing example (cont)

```
   WEIGHT=(0.2038E-03  0.1004E-02  0.4098E-02  0.1392E-01  0.3927E-01
           0.9180E-01  0.1751E+00  0.2552E+00  0.2454E+00  0.1322E+00
           0.3630E-01  0.5034E-02  0.3726E-03  0.1608E-04  0.3201E-07
           0.0000E+00  0.0000E+00  0.0000E+00  0.0000E+00  0.0000E+00);
>QUAD2 POINT=(-0.4064E+01 -0.3636E+01 -0.3209E+01 -0.2781E+01 -0.2353E+01
             -0.1925E+01 -0.1497E+01 -0.1069E+01 -0.6415E+00 -0.2137E+00
              0.2142E+00  0.6420E+00  0.1070E+01  0.1498E+01  0.1926E+01
              0.2353E+01  0.2781E+01  0.3209E+01  0.3637E+01  0.4065E+01),
   WEIGHT=(0.0000E+00  0.0000E+00  0.2593E-06  0.8372E-05  0.1342E-03
           0.1510E-02  0.1186E-01  0.5947E-01  0.1716E+00  0.2743E+00
           0.2553E+00  0.1495E+00  0.5766E-01  0.1527E-01  0.2959E-02
           0.4537E-03  0.4634E-04  0.3569E-05  0.0000E+00  0.0000E+00);
>QUAD3 POINT=(-0.4064E+01 -0.3636E+01 -0.3209E+01 -0.2781E+01 -0.2353E+01
             -0.1925E+01 -0.1497E+01 -0.1069E+01 -0.6415E+00 -0.2137E+00
              0.2142E+00  0.6420E+00  0.1070E+01  0.1498E+01  0.1926E+01
              0.2353E+01  0.2781E+01  0.3209E+01  0.3637E+01  0.4065E+01),
      WEIGHT=(0.0000E+00  0.0000E+00  0.0000E+00  0.0000E+00  0.0000E+00
              0.3837E-06  0.1858E-04  0.4446E-03  0.5815E-02  0.3884E-01
              0.1307E+00  0.2343E+00  0.2498E+00  0.1801E+00  0.9737E-01
              0.4201E-01  0.1487E-01  0.4353E-02  0.1039E-02  0.2166E-03);
>SCORE   IDIST=3,METHOD=2,NOPRINT,INFO=1,POP;
```

BILOG-MG command file for Prototype 2 analysis

```
ANALYSIS 2:  A SIMULATED TWO-STAGE TESTING APPLICATION
  Method 2:  Incomplete block two-stage test
>COMMENT
        Based on the 100-word spelling test data. N = 1000

>GLOBAL  DFNAME='SPELGRPF.DAT',NPARM=2, SAVE;
>SAVE    SCORE='SPEL2N2F.SCO',PARM='SPEL2N2F.PAR';
>LENGTH  NITEMS=36;
>INPUT   NTOT=100,SAMPLE=1000,NFORMS=9,NGROUP=3,KFNAME='SPELGRPF.DAT',
        NIDCH=11,TYPE=1;
>ITEMS   INUM=(1(1)100), INAME=(SPELL001(1)SPELL100);
>TEST    TNAME=SPELLING,INUM=(1,4,5,6,8,9,10,12,14,15,17,23,24,25,
26,27,28,33,34,35,38,47,48,49,50,53,54,60,68,69,73,78,85,90,92,95);
>FORMA1  LENGTH=20,INUM=(5,6,9,14,17,24,25,26,27,38,48,54,68,69,53,78,
85,90,92,95);
>FORMB1  LENGTH=20,INUM=(1,5,8,14,15,24,26,28,33,35,38,49,53,68,73,78,
85,90,92,95);
>FORMC1  LENGTH=20,
INUM=(4,5,10,12,14,23,24,26,34,38,47,50,53,60,68,78, 85,90,92,95);
>FORMA2  LENGTH=20,
INUM=(1,4,5,6,9,10,15,17,25,26,27,33,34,48,49,50,54, 68,69,90);
>FORMB2  LENGTH=20, INUM=(1,4,8,9,10,14,15,27,28,33,34,35,38,48,49,50,
54,73,78,92);
>FORMC2  LENGTH=20, INUM=(1,4,9,10,12,15,23,24,27,33,34,47,48,49,50,
53,54,60,85,95);
>FORMA3  LENGTH=20, INUM=(5,6,8,9,12,17,23,25,26,27,28,35,47,48,54,
60,68,69,73,90);
>FORMB3  LENGTH=20, INUM=(1,6,8,12,14,15,17,23,25,28,33,35,38,47,49,
60,69,73,78,92);
>FORMC3  LENGTH=20, INUM=(4,6,8,10,12,17,23,24,25,28,34,35,47,50,
53,60,69,73,85,95);
>GROUP1  GNAME=GROUP1, LENGTH=36, INUM=(1,4,5,6,8,9,10,12,14,15,
17,23,24,25,26,27,28,33,34,35,38,47,48,49,50,53,54,60,68,69,73,
78,85,90,92,95);
>GROUP2  GNAME=GROUP2,
LENGTH=36,INUM=(1,4,5,6,8,9,10,12,14,15,17,23,24,25,26,27,28,33,34,35,38,47,48,49,50,53,54,60,68,69,73,78,85,90,92,95);

>GROUP3  GNAME=GROUP3, LENGTH=36, INUM=(1,4,5,6,8,9,10,12,14,15,17,
23,24,25,26,27,28,33,34,35,38,47,48,49,50,53,54,60,68,69,73,78,85,90,
92,95);
 (11A1,1X,I1,1X,I1,1X,20A1)
>CALIB   EMPIRICAL,NOFLOAT,CYCLE=35,SPRIOR,NEWTON=2,
        CRIT=0.001 REF=0,PLOT=1.0,ACC=0.0;
>SCORE   IDIST=3,METHOD=2,NOPRINT,INFO=1,POP;
```