DUPLEX DESIGN: GIVING STUDENTS A STAKE IN EDUCATIONAL DEVELOPMENT

CSE Technical Report 306

R. Darrell Bock
Michele F. Zimowski

NORC

UCLA Center for Research on Evaluation,
Standards, and Student Testing

January, 1990

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

PREFACE

Over the past twenty-five years, large-scale educational assessment has grown from a small privately funded endeavor into a major activity of federal and state governments. In the United States, the National Assessment of Educational Progress (NAEP) has periodically surveyed student attainment in main subject-matter areas and reported at a national and regional level since 1969. In 1990, the NAEP sample will be supplemented to allow comparisons between states. Concurrently, all but a few of the states have developed their own annual testing programs to meet local and state needs for information on learning outcomes. On the international stage, some of the Canadian provinces and a number of other countries are planning assessments or have them in place. England and Wales will begin development of a national testing program in 1990. Spain and West Germany have held conferences to explore possibilities for large-scale testing programs for their school systems.

Surprisingly, this remarkable growth of assessment has occurred without much systematic study of the measurement and statistical methodology that supports these programs. Thus far, there are no monographs, and only scattered journal articles, that address the technical problems of assessment. From the little literature that exists describing early plans for a national assessment, one gets the impression that the existing theories of survey sampling and educational measurement were considered sufficient for successful implementation of the program. A number of papers from that period inquired how student learning could be tested using more relevant and revealing tasks than the conventional multiple-choice items, but there was little if any attention to the question of how the assessment could provide consistent measurement over years and decades. The goal of obtaining dependable information for guiding educational policy was clear, but the means were not. The early advocates of assessment made the case for a system for long-term monitoring of student attainment and left the details to be worked out later. And indeed the details were worked out, but mostly ad hoc, and with insufficient prior study or subsequent evaluation. As a result, we are now in a position of having assessment as an accomplished fact, but with little conception of whether it is optimally designed, cost-effective, and accomplishing its goal of guiding pre-collegiate educational policy.

The present report is an attempt to step back from the programs as now constituted and to suggest how assessment instruments and procedures might look if designed as part of a comprehensive educational information system. The work we report includes the actual construction, field testing, and statistical analysis of a prototypical "ideal" assessment instrument. With the support of the United States Department of Education Office of Educational Research and Improvement (OERI), coordinated by the Center for Research on Evaluation, Standards, and Student Testing (CRESST), we have created, conducted, and documented a small-scale model assessment in 8th-grade mathematics. In the three years of the study, we developed a 24-form assessment instrument, and tested it in 32 Illinois public schools; then revised the instrument and tested it again in another 32 schools in California. All of the field work, including the reporting of results to the school and to the state, were carried out by the National Opinion Research Center (NORC) in a way that closely simulated an operational state assessment.

The conception of assessment in this report grew out of many experiences of the first author in educational measurement and assessment projects and committees from about 1960 to the present. These include the Emergency School Aid Study, the Textbook Comparison of the School Math and Study Group, the Analysis Advisory Committee of the National Assessment of Educational Progress, the Technical Advisory Committee of the California Assessment Program, a similar advisory committee to the Illinois Assessment, a Committee of the National Academy of Sciences reviewing the United States contribution to the Second International Mathematics Study, a technical resource committee for the Alexander-James report on the National Assessment, and most recently, the Committee of the National Center for Educational Statistics to review plans for the state-by-state NAEP

The study could not have been carried out without the generous cooperation of the Office of Evaluation of the Illinois State Board of Education, Tom Kerins, Director, and the California Assessment Program, Dale Carlson, Director. The help of Ted Sanders, Superintendent of Education in Illinois, and Bill Honig, Superintendent of

Schools in California, in recruiting the schools is gratefully acknowledged.

We also extend our thanks to many colleagues and co-workers, including Tej Pandy, Mervin Brennan, John Dossey, Judy Wells, Ken Travers, Zalman Usiskin, Joan Herman, Robert Mislevy, Richard Hill, Eiji Muraki, Steve Schilling, and Eve Weinberg. Special thanks is due John Dossey, Judy Wells, and their students, who wrote most of the new mathematics items that were required in the study. John Dossey and Doris Redfield reviewed and suggested improvements in chapters 1 through 4. Margarette Wafer, Peggy Chandler, Steve Schilling and Daniel Cantor assisted in the preparation of camera copy for the report.

The present version is an OERI/CRESST report. A publication version including more detailed statistical results and an additional chapter devoted to procedures for linking state and national assessment scales is in progress.

Contents

Chapter 1

Assessment for Whom?

The concept of educational assessment, and the term itself, originated in the 1960's with Ralph Tyler's initiative to establish a system for monitoring nation-wide outcomes of primary and secondary education. His efforts, and those of his colleagues, were successful in organizing the National Assessment of Educational Progress (NAEP) and eventually securing a Congressional mandate for its continued existence. The program, goals, and methods of NAEP soon came to define a new form of large-scale educational evaluation. Employing matrix-sampling methods for the selection of schools and students to represent national regions and the selection of tasks and exercises to represent subject-matter domains, it was able to obtain at reasonable cost the statistics that could have described trends in educational outcomes over years and decades.

During this period when NAEP was becoming established as a national institution, state and local education agencies were coming under pressure to justify their expenditures of public funds by demonstrating the effectiveness of their schools. The movement for *accountability* led an increasing number of states to establish annual testing programs as a way of measuring the outcomes of their instructional programs. This trend has continued until, according to the most recent survey of the Association of State Assessment Programs, all but two of the fifty states have some form of state-wide standardized testing of student attainment (Roeber, 1988).

Only ten of these programs are policy oriented "assessments" along the lines of NAEP, however. The others may share some of the goals of NAEP, but their methods are those of norm-referenced achievement testing programs or criterion-referenced, minimum competency programs. While these more traditional programs can generate state-level statistics, their primary role is in student guidance, placement, certification, and remediation. They are not specifically designed as information systems for the measurement of the long-term performance of schools but are more responsive to local needs.

In contrast, the assessment design discussed in this report incorporates both the goals and methods of NAEP as they apply at the state level, and those of providing data that more directly serve the needs of the schools and local communities. The design is in harmony with the premise of the accountability movement that public officials have an obligation to expend funds in support of schools and school programs in a way that returns fair value to society. Its purpose is to maximize benefits of assessment to all parties to public education, while minimizing testing time diverted from classroom instruction.

A corollary of the accountability premise is that state initiatives to improve public education must also provide for evaluating the impact of the innovation. Despite the theory and study that may have gone into the planning of a new program or facility, there can be no guarantee that the results will be uniformly favorable. At some point, the state must produce tangible evidence to satisfy the cost-benefit requirement. Rather than rely on local or ad hoc studies for this purpose, many states have moved toward permanent state-wide systems for evaluating student performance in the public schools. In some respects the resulting data are just another form of social statistics (such as rates of unemployment, incidence of communicable diseases, highway accident rates, *etc.*) that state agencies routinely monitor and report. They have the same potential to inform policy and resource planning but are technically more difficult to analyze and report. Administrative statistics such as class sizes, drop-out and graduation rates, teacher qualifications, *etc.*, are relatively easy to assemble and summarize, but they are only indirect symptoms of educational problems. More detailed diagnostic information is available only in direct assessment of student attainments. The present report is concerned with the question of how to obtain such information more effectively than has been possible with conventional assessment or achievement testing methods.

1.1 More favorable cost-benefit for educational assessment

The costs of the assessment itself must, of course, be considered along with the benefits. The costs can be considerable, depending on the extent of the program. A minimal program based on sampling schools, sampling students within schools, and testing at perhaps three grade levels—four, eight, and twelve—is relatively inexpensive in comparison with total state expenditures for education. Increasingly, however, legislatures are mandating detailed assessment reports for every public school in the state; they are requiring a complete census of students in the selected grades, or in some cases, every grade. In a large state, the cost of these more ambitious programs runs to many millions of dollars per annum. The allocation of this money to testing when it could otherwise directly support instruction can be justified only if the assessment commensurately benefits the student and the community.

On the assumption that most state agencies, including the testing programs, are already under pressure to make every dollar count, there is not much prospect of increasing the cost-benefit of educational assessment by reductions on the cost side. The real opportunities for improvement lie in extending the benefits of assessment to broader constituencies within society. The theme of the present study is that, rather than limiting its function to the political and policy uses, the assessment can increase its utility with modest marginal cost by making its data relevant to the planning, management, and conduct of education at the district, school, and ultimately the classroom and student level.

1.2 School-level reporting

A number of states have taken a first step in the direction of greater utility of assessment data by reporting to the district and school level. Although this step requires a census assessment and is more expensive than sampling schools, it makes the results directly meaningful to every community in the state. Moreover, it better serves state policy-makers because the sources of problems in public education are more accurately identified in data for individual schools than in aggregate results for the state. The advantages of having complete information for schools have moved states that previously employed sampling methods, such as Illinois and Pennsylvania, to adopt census assessment and school-level reporting.

The cost considerations that have dictated survey assessments are based on NAEP's practice of sending interviewers to the sampled schools to administer the tests. In a census assessment of all schools, travel and personnel costs of this procedure would be all but prohibitive. The only feasible approach is to use local school personnel, specially trained for the purpose, to administer the tests and return the test materials to the program or contractor. Of course, clear guidelines are required to guard the security of the test forms and to maintain uniform testing conditions, but the experience of states such as California shows that they are understood and observed. Local administration also has the advantage that the test administrator is known to the students and generally able to conduct the testing with less disruption of classroom routine.

To encourage the cooperation of the schools in a locally administered assessment, the program should provide reports and interpretational aids that are relevant and meaningful to principals, teachers, and counselors. For primary and secondary schools, reports at the classroom level as well as the school level are helpful. Features of the data analysis that make such reporting possible should therefore be part of the assessment design. We will have more to say about effective communication of assessment results in Chapter 7.

1.3 Student-level reporting

Assuming that the utility of assessment has been broadened by school-level reporting, the next logical step is to expand the user-community still further by reporting some part of the results at the *student level.* In this way, the assessment reaches directly to the largest constituency of all—the teachers, parents, and students. To accomplish this goal, while retaining the advantages and efficiencies of group-level assessment, we propose the use of measurement methods that provide reasonably reliable scores for individual students, along with efficient measures of program outcomes at the school, district, and state level.

The approach is based on work of Bock and Mislevy (1988) reported in *Educational Evaluation and Policy Analysis*; they introduced a structured form of assessment instrument, called a "Duplex Design", which provides scores in broad content and process categories at the student level jointly with measurement of detailed curricular objectives at the school level. In the present study, we construct and field-test a Duplex Design for eighth-grade mathematics. Our objective is to appraise the potential of the design to serve a wide community of assessment users as well as to improve the quality of the data for present users. The Duplex Design accomplishes this in a way quite different from traditional achievement-testing programs that also report at the student level.

1.4 Assessment programs versus every-student achievement testing

The reason that student-level reporting has not previously been part of educational assessment programs, as inspired by NAEP, lies in the different purposes of traditional every-student achievement testing and policy- or accountability-oriented assessment. The latter programs were originally designed to produce high-level aggregate statistics for policy purposes by using sample-survey technology. They were necessarily top-down affairs not directed toward the needs of classroom teachers, or of parents and students. Indeed, the students have had no part in the process other than that of responding to the assessment exercises. In this situation, it is generally accepted that students must be told that their performance will have no effect on their course grades—that nothing will be reported back to the classroom teacher or to their parents. Students are asked to do their best in responding to the assessment tasks, but are given no personal stake in the results and receive no direct benefit in return for their efforts.

In contrast, traditional achievement-testing programs, which historically have been an initiative of local school districts, are high-stakes activities from the point of view of the student. The districts routinely make available test reports, usually prepared by commercial testing services, for the benefit of teachers, parents, students, and in summary form, for local news media. Although some states (Iowa, for example) continue to rely on compilation of these local testing results for their state-level statistics, the advantages of matrix-sampling assessment have moved many other states in that direction. One of the drawbacks of achievement tests, which are designed for making dependable decisions about student placement or promotion, is that they are rather long and time-consuming. To produce a single score sufficiently reliable for individual measurement, a typical achievement test consists of 40 or more items and requires perhaps 30 minutes of testing time. To obtain scores in four or five content areas with these tests thus requires two or more hours of student time—time that must be taken from instructional use.

Although achievement tests usually have good psychometric properties, other features, in addition to their length, limit their usefulness as assessment instruments. First, they are costly to construct because repeated field trials may be necessary to develop items with indices of difficulty and discrimination suitable at a given grade level. For this reason, and also because of the costs of obtaining a large enough sample for establishing national norms, commercially published achievement tests usually exist in relatively few parallel forms, typically four to six. Because of the expense, the tests often are not revised or their item content refreshed as often as they should be to keep them up-to-date and uncompromised.

Second, the relatively small number of forms puts their item content at risk of becoming known to teachers who, perhaps unconsciously, may steer their students toward best strategies for responding to the test items or toward the correct answers to specific items. The effect is even more pronounced if a single form of the test is used over a number of years and is repeatedly administered by the same teacher, which is very often the case. The result is almost inevitably some compromise at the item level that gives schools an advantage relative to the conditions under which the test was normed, before item-exposure could operate. This is perhaps the best explanation for the so-called "Lake Woebegone Effect"—the phenomenon of student performance in all states exceeding the national median as indicated by the no-longer valid norms (Cannell, 1988). A state-developed matrix-sampled assessment instrument based on more forms and items, with provisions for frequent updating of the item content, is far more resistant to these effects.

Commercially published achievement tests can also be problematic for a state testing program because their content is not arrived at by any consensus-making process that reflects current goals of mathematics education in the state. They tend to emphasize computational and procedural skills that are easy to measure in the multiple-choice format, to the neglect of conceptual understanding and strategies of problem solving. For these reasons, state departments or boards of education often prefer to develop their own testing instruments, even though they may not have the resources to develop high-quality achievement tests. In this situation, the scope and efficiency of assessment methodology become very attractive. Because many curricular objectives can be evaluated simultaneously, consensus on the content of the instrument requires fewer compromises and is thus easier to obtain. Finally, as we discuss in more detail below, assessment methods also measure attainment levels in schools, districts, counties and in the state as a whole with better generalizability than is typical of achievement testing. Moreover, because each student responds to fewer items, the booklets of an assessment instrument can contain more varied item formats. Reading passages can be longer, and more intermediate steps in problem-solving exercises can be probed. Considerations such as these argue persuasively for the choice of assessment designs over achievement testing as the preferred method of monitoring educational attainment in the public schools.

1.5 Matrix-sampling assessment designs

Matrix-sampling applications in education had their origin in the realization that precise estimation at the individual level is not necessary for accurate estimation of average attainment of *groups of students*. By suitable statistical methods, it is possible to estimate mean scores for schools, school districts, instructional programs, demographic groups, or the state as a whole, directly from the item responses. If the group-level scores are estimated from the responses of a large number of students, the limiting factor in the accuracy of the result is not the reliability of individual measurement, but rather the generalizability to the content domain. The somewhat surprising conclusion is that the required generalizability can be obtained by administering sufficient numbers of items from the domain, *each to a different student.*

When Frederic Lord, at the suggestion of William Turnbull, examined the possibilities of this approach to group measurement, he found that the most efficient design was one in which each student is assigned, at random, exactly one item from the domain sample (Lord, 1962). If the students are also regarded as a sample, this technique of conducting a survey is called "matrix sampling". The population of persons is represented by the rows of the matrix and the item domain by the columns. The sample obtained by randomly selecting rows and columns and obtaining the response of the corresponding person to the corresponding item constitutes the *matrix sample.* If the persons are assigned items in this way from several distinct content domains, the procedure is referred to as *multiple-matrix sampling.* The statistical properties of multiple-matrix sampling have been extensively studied from the classical point of view by Serotnik & Wellington (1977), and from the point of view of item-response theory by Bock and Mislevy (1981).

The application of multiple-matrix sampling to educational assessment leads to assessment instruments consisting of many forms, 20 to 40, each consisting of perhaps 40 to 50 items. When the responses of a group of students are aggregated across forms, the instrument can measure as many distinct objectives as there are items in each form. This is about the maximum number that teachers can profitably use. If for some reason curricular specialists wish to evaluate a greater number of objectives, the number can be doubled by creating two such instruments and assigning them in alternating rotation within classrooms.

The principal advantage of matrix-sampling designs is the high level of generalizability they provide with relatively small demand on classroom time. The higher the generalizability, the more dependable is the instrument in characterizing the average attainment levels of the schools. Matrix-sampling designs are able to provide this dependability for the following reason. Even slight variation in instructional programs or emphasis will result in interactions between the schools and particular items that are sensitive to such differences. Because the items are randomly selected to make up an assessment instrument, these interactions constitute errors of measurement, relative to some other set of items randomly selected from the same domain. The independence of the item-by-school interactions under random item selection assures, however, that the size of the error variation in the school score will diminish proportionally as the number of items increases. At the same time, the variability of the school mean-score is also affected by number of students tested; the larger the school, the more precise is the mean score.

These effects are apparent in the year-to-year correlations for California public schools as shown in Table 1.1. The data are reading-score school means measured in two successive years. Note that the sizes of the correlations increase with the number of students tested, and also with the number of items in the assessment instrument. The latter effect is the contribution of item sampling; it would be even greater if different items were sampled each year, because the suppression of the item-by-school interaction that attenuates the correlation would then be more apparent. For this reason, the number of forms in an assessment instrument must be large in order to assure the year-to-year stability of the school-, district-, and state-level scores. Dependable group-level scores will result if large numbers of students take large numbers of different items.

1.5.1 The efficiencies of matrix sampling

Although a large number of forms are required, assessment instruments are in some ways not as difficult to construct as achievement tests. This is partly because each assessment-test booklet has many fewer items than an achievement-test booklet, and also because the difficulty levels of the items do not have to be controlled as carefully as in achievement tests. In assessment, the variation in difficulty of the items and the variation in the ability of students within the school

TABLE 1.1

Effect of sampling of students and sampling of items on the year-to-year correlations of sixth-grade mean reading-attainment scores of California schools

		Number of items in matrix sample		
		85	128	400
Number of students	50	.59	.73	.79
sampled	100	.67	.78	.88
per grade	200	.76	.81	.93

combine to prevent the item percent-correct statistics from becoming too extreme and thus impairing the functioning of the instrument. Although it might be thought that wide differences between public schools within a state could lead to situations where all or none of the students succeed on some of the items, this seldom happens in practice. Within-school variation is always considerably larger than the between-school variation and guarantees variation in the observed responses. We give some estimates of within- and between-school variances for California schools in Chapter 6.

Typically, an assessment form is about one-third the length of an achievement test. Thus, the number of distinct items required for a thirty-form assessment instrument would be the same as ten forms of an achievement test. But the economies of scale in producing and selecting items reduces the effort in constructing an assessment instrument to about the same order of magnitude as producing perhaps five or six forms of an achievement test.

1.5.2 The hidden "costs" of matrix-sampling designs

Regrettably, the efficiencies of multiple-matrix sampling designs carry with them certain less tangible costs. The greatest of these is their failure to provide useful information about individual students. They do not permit an individual student's scores to be reported to classroom teachers, to parents, or to the students themselves. Thus, the students have no personal stake in the outcome of the assessment testing apart from their loyalty to the school, the strength of which may vary considerably from one school to another.

We submit that the use of unmotivated testing is a questionable practice from the point of view of both measurement and pedagogy. Measurement of scholastic performance necessarily requires a best effort on the part of the student—otherwise, the student's score has no definite meaning. This may be understood by analogy with the measurement of the height of a mountain: unless made at the highest point, the measurement is not a unique value—any other elevation would be quite arbitrary. Similarly, the student not doing his or her best work is not displaying any definite or easily reproducible level of proficiency that a test can measure. Under conditions of variable motivation, test scores are not comparable and stable.

This is not the case when the student is personally involved with the outcome. For example, when students retake the SAT to improve their chances for college admission, they seldom change their scores by more than 10 or 15 points in a range of 600. The results tend to be consistent because optimum performance is a more stable individual attribute than the situationally dependent performance of less-motivated respondents.

There is evidence that poorly motivated testing can have serious consequences when tests are developed or standardized in samples of only marginally interested volunteers but are applied in a high-stakes testing program. A recent example is the experience with the Texas in-service teacher qualification test. Shepard and Kreitzer (1987) have reported that, whereas 12 percent of the volunteer field-trial sample for the test scored below the assigned passing level, only 3.3 percent of teachers failed the test when it was administered operationally.

Similarly, Schilling and Bock (1989) found that scores on the Degrees of Reading Power (DRP) test administered to all twelfth-grade students in California averaged substantially below the national DRP norm. Their performance on this test was considerably below that of the CAP overall reading score when expressed on the NAEP nationally-normed reading scale. The difference is that the CAP reading assessment is an integral part of the annual California school evaluations, whereas the DRP was administered experimentally with instructions that the scores do not count and will not be reported. In contrast, the DRP national norms are based on demographically adjusted student scores in the State of New York where satisfactory performance on the test is required for graduation. These examples illustrate the fact, long accepted in the field of educational measurement, that between-group comparisons of attainment-test results are possible only when the motivational conditions are comparable.

The great danger that untoward motivational effects pose for state statistics on attainment is that they will interact with background variables such as race, social class, education of parents, or school program. Academically oriented students are inclined to accept any test as a challenge and do their best; the same performance is much less likely from a student who takes little interest in intellectual pursuits. Lacking personal involvement, the latter may not respond at a level that is indicative of his or her true capacity.

The case that unmotivated testing is poor pedagogy is even more clear. Current research is reasserting the importance of student involvement and active participation in learning, the recognition of which goes back to Dewey and beyond (see Stodolsky, 1988). To present the student with a test that is unrelated to classroom activity, and worse, to provide no report of the result, utterly violates this principle. Testing under these conditions is just one more exercise that fails to engage the student actively in a self-sustaining process of learning. It violates the accepted motivational principle of feedback of formative information to the student (Brophy, 1987).

Parents could also perceive such testing as an unwarranted diversion of classroom time from instruction or other more constructive use. The time demand will be an especially sensitive point if the school also employs a commercial testing service to provide measurement of individual-student attainment for guidance purposes and reports to parents. Although assessment testing normally requires only one class period while achievement testing may require two or three, the additional time may seem unwarranted considering that the item content of the two procedures is highly similar. Many students, especially those who are older or more perceptive, will view the assessment exercise as having no meaning or consequences for them and will take it as evidence of the artificiality of the instructional program. Thus, the efficiencies of matrix sampling may be purchased at some cost to student morale.

Ironically, the lack of individual-student scores in the matrix-sampling design also has unfavorable consequences for the very clientele it is designed to serve. Although these designs enable public officials to point to group-level statistics such as average-percent-correct or scale scores as indicators of the relative state of educational attainment, they do not allow the kind of terminology that is most natural to public discussion— namely, the citing of numbers or percentages of students who are reaching a satisfactory level of attainment in pass-fail terms. The difficulty is that these terms require information about the *distribution* of levels of attainment in the population of students. If accurate estimates of the attainment levels of individual students are available, these percentages can be obtained simply by counting the number of students who exceed the criterion level and dividing by the total number of students. If the scores for the students are measured with appreciable error, more complex statistical methods are required to estimate the required percentage. In either case, the required information is available only if the assessment instrument provides scores for individual students, which the matrix designs do not.

1.6 The Duplex Design

The Duplex Design combines in one instrument the functions of individual-student achievement testing in broad content areas and those of group-level assessment of detailed curricular objectives. The term "Duplex Design" refers to the two-fold manner in which the original data—the item responses—are used to obtain these two types of information from a single test administration.

Basically, the Duplex Design is a replication of a content-by-process item arrangement in a number of test booklets. A schematic representation of this arrangement appears in Figure 1.1.

The information extracted from the test booklets is further enhanced by the use of conjoint scoring to estimate the content and

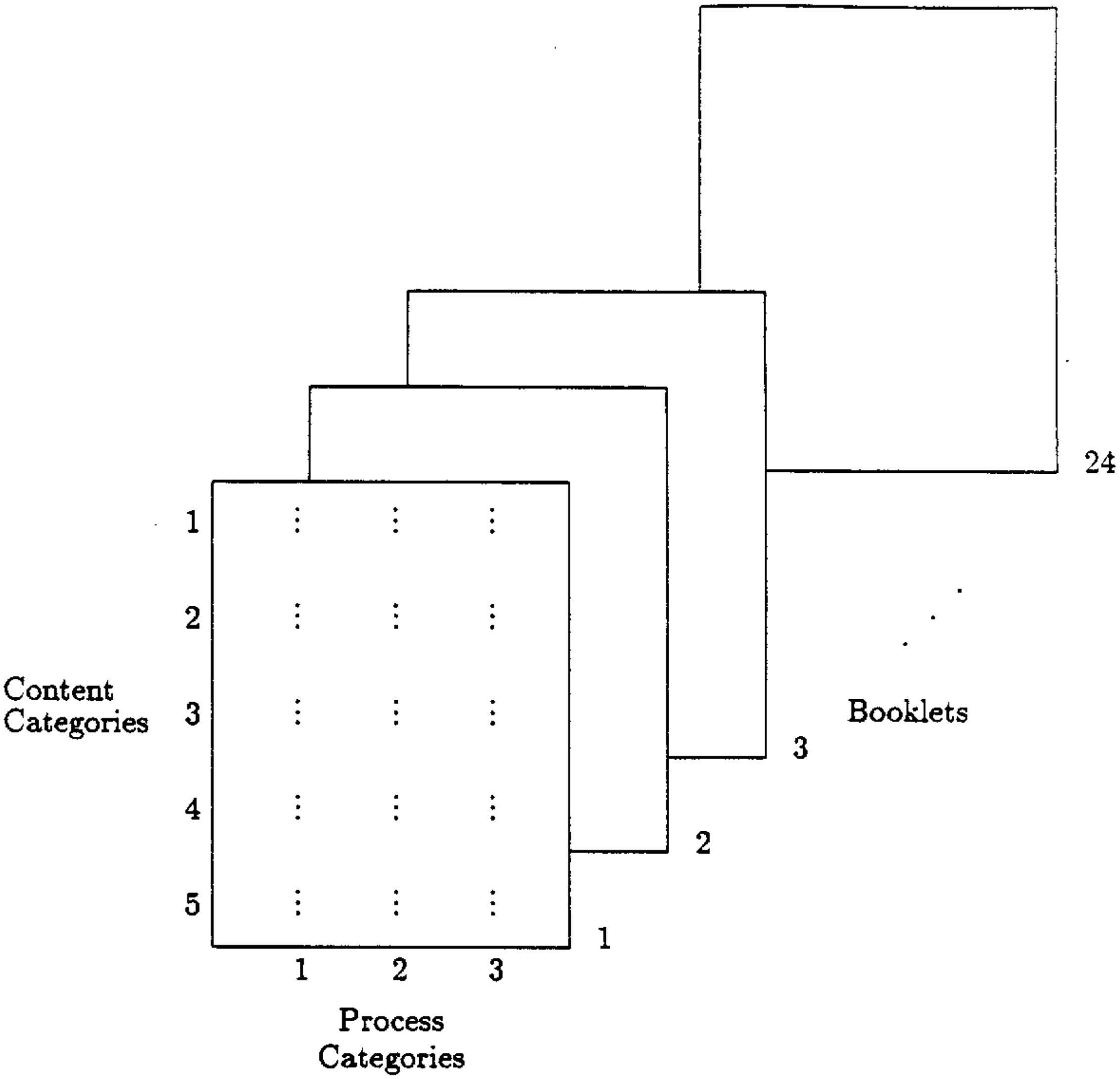


Figure 1.1. Layout of a typical Duplex Design. Topics within content categories are replicated in each booklet. Items classified by content categories and process categories are randomly assigned to forms.

process scores from responses to the same items. Because each item is classified by its content and process dimension, the item structure within each booklet is formally identical to a two-factor experimental design in which the items are arranged in a row-by-column table. In the arrangement shown in Figure 1.1, the rows correspond to content categories, and the columns to process categories. The scoring methods discussed in Chapter 3 are used to aggregate the item responses across rows to obtain the *content scores*, and down the columns to obtain the *process-proficiency scores*.

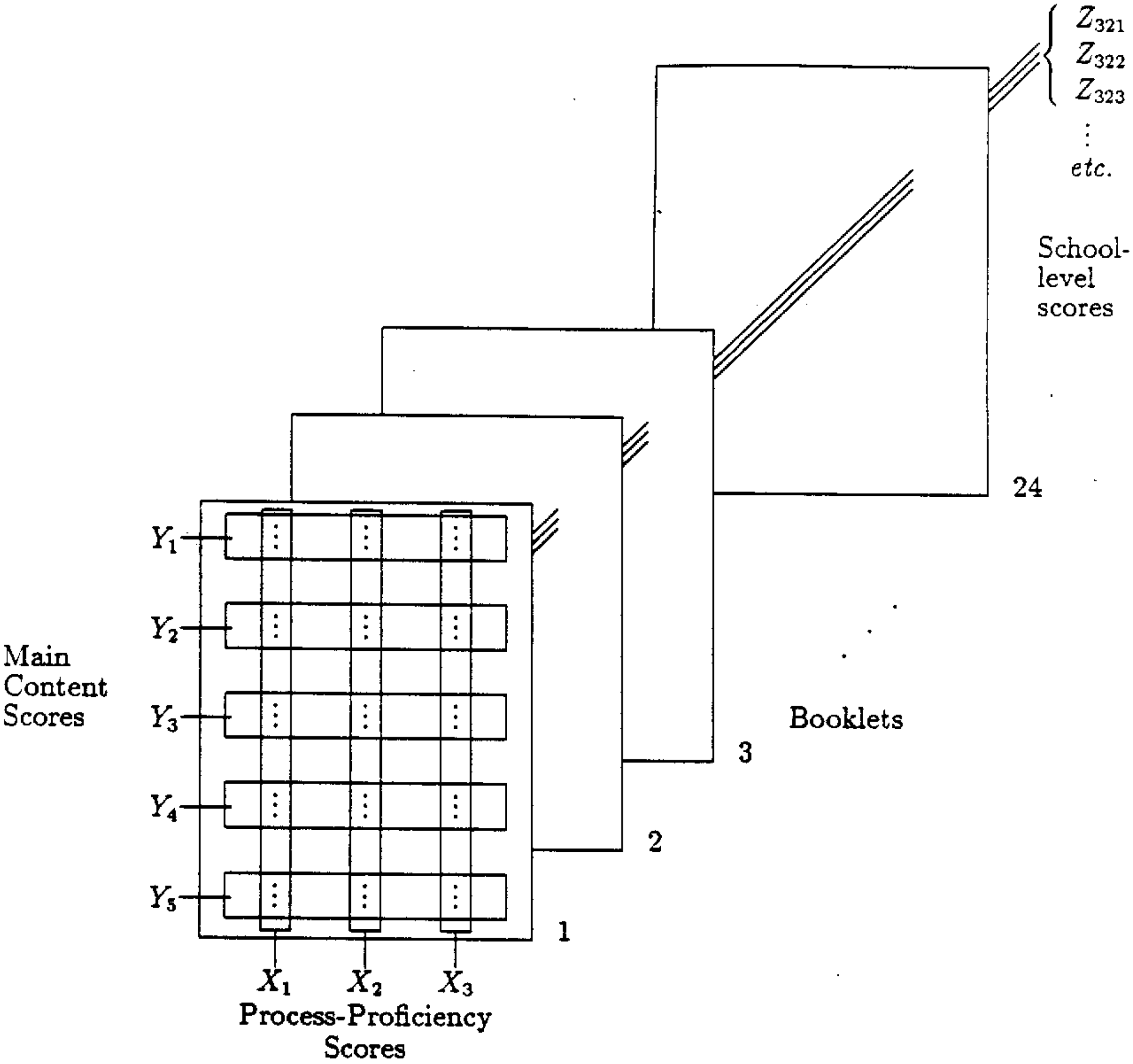
These student-level scores can then be readily aggregated to various group levels—the classroom, the school, the district, county, and state. At the same time, items in each *cell* of the content-by-process classification can be aggregated, by methods described in Chapter 3, across the test booklets in order to obtain a *group-level score* for the corresponding curricular objective. In this way, as many curricular objectives as there are items on the test can be measured at the group level. Figure 1.2 represents these types of scoring of the Duplex Design. Further details of the Duplex Design principles and analysis are discussed in Chapters 2 and 3.

1.7 Uses of group-level and student-level assessment data

Potential uses of state assessment results have been review by Cohen (1988) and Bock & Mislevy (1988). Some of these uses need only school-, district-, or state-level data; others require the scores for individual students that the Duplex Design provides.

1.7.1 Policy formulation

Supplying information to guide and justify state educational policy is the assessment's most visible role. If new programs affecting such factors as teacher recruitment, time and resources devoted to instruction, changes in standards of grading and promotion, and school administration are to be adopted and institutionalized, public officials need evidence that these reforms are really improving student attainment. An assessment that delivers comparable and dependable year-to-year



Subject-level scores

Figure 1.2. Scoring the Duplex Design. A student responding to one of the booklets receives a score on the main content and process categories. A school receives a score on each topic within cells of the content-by-process classification.

data can provide such evidence in a space of two or three years if a suitable base line has been established prior to the change in policy. Obviously, the capability of the assessment instrument to produce comparable and accurate measures of attainment in the main areas of instruction is essential in this role. The validity of the assessment as a long-term monitor of educational progress requires a stable and conservative organizational base, which presumably will reside in the state agency that develops and implements the assessment. That agency will have the responsibility for insuring that a consistent methodology is maintained as personnel, contractors, and facilities change.

The assessment can also have a role in initiating and sustaining political support of educational reform. State officials are conscious of the impact of the quality of public schools on economic development. Good schools make it easier to attract capable workers to the state and thus provide the personnel for commercial and industrial development. If the assessment results are good, they can justify existing budgets or even the expenditure of more funds to build on strengths. If the results are poor, they can generate support for efforts to improve facilities and instruction. Without assessment data, political leaders have few facts on which to make the case for educational improvement.

1.7.2 Media reporting

The news media routinely give prominent coverage to educational assessment reports, usually with some attempt at interpretation. Reporters are accustomed to discussing statistical indices, such as the Dow-Jones index of stock prices, that are on an arbitrary scale not unlike a standardized educational attainment score. Usually they use these indices to make some sort of comparisons, either in the form of a graph showing gain or loss from previous years, or a bar chart showing differences between groups. From properly prepared data, they can also make many within-state comparisons—between school districts, cities, regions, ethnic groups, *etc.*

The media will also have a great interest in comparing state results with those of other states or the nation. Unfortunately, the necessary figures are not normally available for an assessment instrument developed by the state, and the opportunities of obtaining such information by having such instruments administered in other states are limited. A better scheme is to relate the state instrument to an already existing nationally-normed test so that the main state results can be expressed on the scale of the national test. The state assessment agency can then supply accurate comparisons of state and national results to the media.

The state assessment agency can establish the relationships between their instrument and the national test by arranging for a sample of students who will routinely take the state test to also take the national test. Because both tests are operational, motivation effects are consistent. The relationship between the two instruments can then be determined statistically from the paired scores of these students. The two instruments should, of course, be as similar in content as possible, but they need not be identical. Nor do the two tests have to be taken during the same school term. Provided the state test can accurately predict scores on the national test, the results of the state assessment are validly expressed on the nationally-normed scale. Schilling & Bock (1989) have demonstrated such a procedure by expressing reading and mathematics results from the California Assessment Program in terms of the corresponding scales of the National Assessment of Educational Progress.

1.7.3 Educational planning

State departments of education or similar agencies can benefit from the assessment in ways more specifically related to instruction. If the assessment measures success in attaining detailed curricular objectives, and includes breakdowns by county, district, and relevant demographic and economic factors, the agency will have a clear picture of the strengths and weaknesses of the state's educational program. Comparison of these data with teaching practices in the schools, choice of textbooks, student motivation, and community involvement, *etc.*, may suggest new strategies for shaping and strengthening the school system. Where problems exist, the school-level scores will provide diagnostic information for tracing the difficulty to such sources as curricular emphasis or timing, student-grouping policies, instructional strategies, teacher qualifications and preparation, *etc.*

Some states have used assessment data to direct additional public funds to school districts with obvious deficiencies in attainment; others have taken the opposite tack and provided monetary rewards for schools with exceptionally high performance relative to what could be expected from their community background and resources. We discuss in Chapter 7 a statistical method of accounting for background characteristics of the school by means of so-called "comparison score bands" devised by Cronbach for the California Assessment Program.

1.7.4 School management

At the district and school level, the wealth of detail from a matrix-sampling design provides instructional effects specific to programs, courses, and even units within courses. Here, the overall level of attainment in the school is not as important as the profile of strengths and weaknesses over the various process and content categories of the design. As we show in Chapter 7, this information can be displayed in a graphic form that serves well for group discussion among members of the teaching staff. In this use, the great advantage of matrix-sampling assessment at the school level is apparent. Traditional achievement-test profiles are so general as to support only vague exhortations for a better performance, while matrix-sampling designs can report to the school directly in the topics and skills that appear in course outlines. Teachers can then look for reasons for the detailed outcomes in so far as their students influence the overall school performance. In large secondary schools where teachers have more than one section of the same subject, it may be possible to report the assessment groups broken out by teacher. Provided the varying input of students assigned to classrooms is accounted for, a "between-teacher" analysis can reveal interesting effects of teaching styles on attainment outcomes. We present some analyses of this kind in Chapter 7.

1.7.5 Counseling and guidance

Counseling of individual students obviously requires student data with some degree of diagnostic detail. The Duplex Design provides this information in a score profile that indicates the student's strengths and weaknesses in main subject-matter areas and in general skills. These profiles are useful in conferences between counselors and parents to make clear where the student's achievement problems may lie. In this role, the assessment results have the advantage of objectivity and impartiality, and they can be referred to the local and state norms to convey how well the student is performing relative to his or her peers. In Chapter 7 we illustrate graphical devices for presenting this information in an easily understood form.

Classroom teachers can also benefit from such information if a copy of the individual-score profile is provided at the time class assignments are made. Typically, the assessment is administered late in the school year and results are prepared over the summer in time for the opening of a new school term. This is an appropriate time for computerized student reports to be supplied to the classroom or homeroom teacher for the new term. At the same time, these reports can be distributed to the students and passed on to parents. Although purely advisory, they will carry the prestige of the state assessment program and provide a useful basis for guidance conferences early in the term. By computerized methods, the reports can include not only a numerical and graphical presentation of the students progress in the previous school year, but can also contain a computer-generated verbal interpretation of the individual results. This form of reporting involves the parent and the student in the assessment in an immediate way, and gives the student a personal stake in its outcome. More than any other factor, it is this enhancement in motivation and personal relevance that recommends a Duplex assessment design capable of reporting at the student level.

1.7.6 Secondary uses of assessment data

Student-level reporting also improves the cost/benefit of the assessment program by producing data that is more suitable for secondary research than are school-level or state-level reports alone. Although it is possible for advanced workers to use both school- and individual-level data, most of the well-known statistical techniques are designed for use with individual-case data. Social-survey analysis as it is presently practiced takes the individual person as the unit of analysis. Conventional matrix-sampling data is not in this form and has therefore not been a productive source for secondary analysis (see Sebring & Baruch, 1983). This has been true not only for state assessments, where secondary analysis has never been emphasized, but also in the national assessment, which was specifically intended to provide a national data base for independent analyses of educational phenomena. Although NAEP data has regularly been made available on user-tapes, the form of the data until recently has made it inaccessible to most research workers. Even the present NAEP users-data consist of multiply imputed scores rather than conventional case-level scores. The new developments in the multi-level analysis of educational data have greatly increased interest in the data bases of the national and state assessments (see Bock, 1989), but the potential of these methods depends on the availability of student-level scores of good technical quality. We demonstrate in Chapter 5 that the Duplex Design can provide such data.

1.8 Another view of assessment: the Assessment-Driven Curriculum

Although assessment was originally conceived as a method of monitoring educational progress, it has recently been seen as a way in which state education agencies can influence the curricula of local schools without directly setting the objectives of instruction. This view is an elaboration of Popham's (1987) concept of "Measurement Driven Instruction", by which teachers are influenced to include and emphasize certain topics because they are routinely evaluated in "high-stakes" state-wide testing.

Assessment can be viewed as a similar way of influencing local curriculum without publishing obligate guidelines at the state level. If the assessment program is the vehicle of accountability, then the content of the assessment instrument defines the student attainments for which schools will be held responsible. The local schools are thus under pressure to teach what is tested by the assessment. The curriculum, which may in principle be the responsibility of the school or school district, becomes "assessment-driven" even if the state provides no curriculum guidelines whatsoever. The content of the assessment instrument sends a clear message about what schools should be teaching.

The concept of an "assessment-driven curriculum" has implications not only for how the assessment instrument is constructed and what content it includes, but more precisely for what scores are reported. School administrators and teachers only see the assessment results in terms of the scores, but there are no broadly accepted views on the amount of detail the assessment scores should reflect. According to the educational philosophy in vogue, some states may confine their attention to broad outcome measures in the main subject-matter areas and allow the schools to decide the kinds of objectives and instruction that are likely to improve scores on these broad measures. Others may believe that local school personnel require the help of curriculum experts in determining more specific content of instruction, and so want the assessment instrument scored for many detailed curricular objectives. One of the attractions of matrix-sampling designs is that they can provide this kind of detailed scoring, at least at the school level, whereas traditional achievement tests, which report only a small number of scores covering main subject-matter areas, are of little value in shaping the local curriculum in detail. Even the most extensive matrix-sampling design has its limitations, however, for if too many objectives are scored, than the number of items per objective may be so small that generalizability will suffer. We find that perhaps 16 items is the minimum number on which a school-level score should be based to provide acceptable generalizability. This means that an assessment instrument consisting of 32 forms each with, say, 45 items, can measure 90 objectives. And even this number is gained at the expense of having only half the students in the school responding to each objective, and in schools with fewer than 100 students the standard error of the school score will suffer (*i.e.*, become excessively large.) Given an amount of testing time limited to one or two class periods, anyone wishing to evaluate curricular objectives in detail has no alternative but to find a middle level of generality where the number of distinct objectives to be scored does not exceed something on the order of 50 to 100 distinct topics.

1.9 Summary

State-wide testing programs are now established components of public education in many states and are in advanced stages of implementation in others. To realize their full potential and to justify their costs, these programs should serve the widest possible community of users. In this, assessment programs that test all students at several benchmark grade-levels and report to every school have the advantage over more limited programs that sample schools and report only at the state level. To be even more beneficial, the assessment instrument should be designed in such a way that scores in broad content areas can be reported to students, while much more detailed evaluations of curricular objectives can be reported to the schools and aggregated to the district and state level.

The Duplex Design is capable of this two-fold type of reporting. It can enhance student motivation, allow results to be discussed in terms of percents-of-students reaching defined attainment standards, and support reliable case-by-case data for secondary analysis. At the student level, the design provides score profiles that can be reported to teachers, parents, and students as aids in student counseling, placement, and certification. At the school level, it provides specific measures of strengths and weaknesses of instruction relative to schools with similar demographic characteristics.

In this report, we describe and evaluate two large-scale field trials of a Duplex Design for eighth-grade mathematics carried out in the states of Illinois and California, respectively. We also discuss the concepts underlying the design and suggest how its results should be reported.

Chapter 2

Principles of the Duplex Design

In common with other assessment methodology, the Duplex Design depends upon matrix sampling to provide generalizable measurement of detailed curricular objectives at the group level without excessive demands on student time. The matrix-sampling principle calls for a test instrument consisting of many forms, each containing an item randomly selected from each of the domains defined by the objectives. Where the Duplex Design departs from standard matrix sampling is in the structure of the item content within each of these forms and in the method of scoring the instrument. The design invokes three further principles to obtain, from the same item responses, within-form score profiles for individual students and across-form measurements of curricular objectives attained by groups or programs:

1. Replication of a content-by-process classification of items within each form. The content and process categories, identical in each form, define the measures represented in the individual student-attainment profiles.

2. Assumption of a multi-level scoring model based on the design structure. The model identifies state-level, school-level, and student-level effects; it includes components attributable to content, to process, and to content-process interaction, plus a residual error.

3. Two-stage testing. Testing time is minimized by an adaptive testing procedure in which a preliminary "routing" test assigns

> each student to a second-stage test booklet tailored to his or her general level of attainment in the subject matter. This type of test administration increases the student's acceptance of the test content and improves the measurement properties of the assessment scores.

Other assumptions are implicit in the Duplex design. It is assumed that student learning can by measured by performance on a suitable number of relatively brief tasks that are easily scored. Although these tasks do not have to be multiple-choice items, they must be limited to sufficiently short answers that the student can respond to some 30 to 50 distinct items during the test period. The presentation of the exercise can be relatively lengthy (for example, a reading passage occupying most of the page of a test booklet), but the items based on the presentation must be varied enough to fill a number of content-by-process categories.

The Duplex concept does not, however, readily extend to essay tests or other exercises requiring complex or lengthy responses. Though it is quite possible to matrix-sample the prompts for essay tests (the California Assessment Program does so in its Direct Writing Assessment), some form of analytic rating of the responses must take place of the content-by-process categories in order to define student-level score profiles and school-level objectives. These extensions of the matrix-sampling methodology are beyond the scope of the present report.

2.1 Principles of instrument construction: the content-by-process classification of assessment tasks

Any scheme to evaluate student learning necessarily inherits the historical lines along which subject-matters are divided for purposes of instruction and study. Typical group-level matrix-sampling designs may cover as many as four subject-matter areas simultaneously. But a group-level *and* student-level Duplex instrument cannot cover more than two subject-matters with enough items per area to allow content and process scoring at the student level. In the present study, for more scope for diagnostic scoring at the student level, we have concentrated on an instrument devoted exclusively to only one subject-matter, mathematics.

2.1.1 The content

By far the most ambitious attempt to reach some agreement on the essential content of school mathematics was the Second International Mathematics Study (SIMS) of the International Association for the Evaluation of Educational Achievement (IEA). In 1979, a curriculum committee of the association with representatives from many nations, nominated the following main content categories for mathematics suitable at a grade level where the modal age is 13 (grade 8 in the United States):

1. Arithmetic
2. Algebra
3. Geometry
4. Descriptive Statistics
5. Measurement

Currently, the same five categories appear in many of the state curricula for eighth-grade mathematics. Within these main categories, however, the topics of instruction enjoy less universal agreement. To the extent that the subject-matter areas coincide with an academic discipline, as in the case of mathematics, the scholarly distinctions in the field have historically influenced the choice of topics, even when their importance for general education is not entirely clear. Partly for this reason, state testing programs seek a broad consensus both in and outside the school system when defining the detailed content of the evaluation instrument. The SIMS took a similar approach in asking teachers and educators from the participating countries to rate the importance of topics in the fairly lengthy list shown in Table 2.1.

To create a prototype Duplex Design for eighth-grade mathematics, we proceeded along similar lines by first combining the relevant

TABLE 2.1

Content categories and topics of the IEA Second International Mathematics Study

0	Arithmetic	
	001	Natural numbers and whole numbers
	002	Common fractions
	003	Decimal fractions
	004	Ratio, proportion, percentage
	005	Number theory
	006	Powers and exponents
	007	Other numeration systems
	008	Square roots
	009	Dimensional Analysis
1	Algebra	
	101	Integers
	102	Rationals
	103	Integer exponents
	104	Formulas and algebraic expressions
	105	Polynomials and rational expressions
	106	Equations and inequations (linear only)
	107	Relations and functions
	108	Systems of linear equations
	109	finite systems
	110	Finite sets
	111	Flowcharts and programming
	112	Real numbers
2	Geometry	
	201	Classification of plane figures
	202	Properties of plane figures
	203	Congruence of plane figures
	204	Similarity of plane figures
	205	Geometric constructions
	206	Pythagorean triangles
	207	Coordinates
	208	Simple deductions
	209	Informal transformations in geometry
	210	Relationships between lines and planes in space
	211	Solids (symmetry properties)
	212	Spatial visualization and representation
	213	Orientation (spatial)
	214	Decomposition of figures
	215	Transformational geometry
3	Probability and statistics	
	301	Data collection
	302	Organization of data
	303	Representation of data
	304	Interpretation of data (mean, median, mode)
	305	Combinatorics
	306	Outcomes, sample spaces and events
	307	Counting of sets, $P(A \cup B)$, $P(A \cap B)$, independent events
	308	Mutually exclusive events
	309	Complementary events
4	Measurement	
	401	Standard units of measure
	402	Estimation
	403	Approximation
	404	Determination of measures: areas, volumes, etc.

curricular guidelines from the assessment programs of the two participating states, Illinois and California. We then turned to a committee of mathematics education specialists to decide on the composition of a common list. The committee consisted of Mervin Brennan, from the Illinois assessment, John Dossey, Professor of Mathematics at the Illinois State University at Normal, Tej Pandy from the California Assessment Program, and Zalman Usiskin, Professor of Education at the University of Chicago. The content categories arrived at by this process appear in Table 2.2.

The main content categories of Table 2.2 agree with those of the SIMS in Table 2.1, except the term "numbers" takes the place of "arithmetic". The subcategories in Table 2.2 are less detailed than those of the SIMS, although in most cases they merely collapse the finer distinctions of that study. The greater detail is, in fact, represented in the item domains from which we selected our items, but none of the finer distinctions appear in the analysis or scoring of the data because the amount of testing time available and the number of forms in the instrument did not permit reporting at so great a level of detail. In addition, we had to omit four of the nineteen subcategories in Table 2.2 from the prototype because we did not have suitable items available for evaluating them (see Table 2.3).

2.1.2 The processes

Recognition of the process dimension in educational evaluation has a much shorter history than that of content. It is linked rather directly to the concept of behavioral objectives of instruction introduced by Ralph Tyler in the early 1950's (Tyler, 1956). According to Tyler, it is not sufficient to define objectives of education merely in terms of *knowing* a topic; rather, the kinds of *behavior* in which knowledge of the topic content is expressed must be made explicit in order to guide instruction. This concept, when applied to educational evaluation, requires a definition of test items that incorporates a behavioral specification as well as a content definition. The result is a schema for evaluation that is in effect a logical product of content and behavior categories. It is conveniently represented by a two-way table in which the rows are the content categories and the columns are the

TABLE 2.2
A grade 8 mathematics Duplex Design

		Proficiencies	
Content Categories	a. Procedural Skills[a]	b. Knowledge of Facts & Concepts[b]	c. Higher Level Thinking[c]
10. *Numbers*			
Integers	11a	11b	11c
Fractions	12a	12b	12c
Percent	13a	13b	13c
Decimals	14a	14b	14c
Irrationals	15a	15b	15c
20. *Algebra*			
Expressions	21a	21b	21c
Equations	22a	22b	22c
Inequalities	23a	23b	23c
Functions	24a	24b	24c
30. *Geometry*			
Figures	31a	31b	31c
Relations & Transformations	32a	32b	32c
Coordinates	33a	33b	33c
40. *Measurement*			
English & metric units	41a	41b	41c
Length, area & volume	42a	42b	42c
Angular measure	43a	43b	43c
Other systems (time, *etc.*)	44a	44b	44c
50. *Probability & Statistics*			
Probability	51a	51b	51c
Experiments & surveys	52a	52b	52c
Descriptive Statistics	53a	53b	53c

[a] Calculating, rewriting, constructing, estimating, executing algorithms.
[b] Terms, definitions, concepts, principles.
[c] Proof, reasoning, problem solving, real-world applications.

behavioral categories.

This schema appeared in Tyler (1956) and later became a rubric for the *Taxonomy of Educational Objectives*, edited by Bloom (1956). Constructed by a committee of college examiners, the Taxonomy identified behavioral objectives suitable for undergraduate courses and considered in detail how to evaluate such objectives using, wherever possible, multiple-choice and short-answer items. Robert Wood (1968) adapted this approach to the school mathematics curriculum in order to classify items for an item-banking project. His behavioral categories were

1. Knowledge and information: recall of definitions, notations, and concepts
2. Techniques and skill: computation, manipulation of symbols
3. Comprehension: capacity to understand problems, to translate symbolic forms, to follow and extend reasoning
4. Application of appropriate concepts in unfamiliar mathematical situations
5. Inventiveness: reasoning creatively in mathematics

The SIMS evaluators used a similar cross-classification of content topics and behavioral categories (the "International Grid") to specify items for the cognitive instrument in the study. Their choice of behavioral categories was,

1. Computation
2. Comprehension
3. Application
4. Analysis

The School Math Study Group (SMSG) used the same categories in its textbook evaluation (Begle and Wilson, 1970).

More recently, with the growing influence of cognitive psychology, evaluation theory has shifted emphasis from the behavioral outcomes to the latent processes involved in task performance—processes that might be inferred from more than one type of overt behavior (see Resnick & Ford, 1981). According to this view, evaluation needs a psychological theory that identifies objective features of tasks that depend on these processes. With such a theory, the test constructor can generate items that manifest these features.

In the field of mathematics education, a number of such theories have begun to appear in the research literature. Skemp (1987), for example, draws a major distinction between *algorithmic* processes of and *conceptual* processes of task solution. He identifies the former with "instrumental" mathematics—a collection of strict, rule-driven procedures for solving standard classes of problems. The latter he calls "relational" mathematics—the creation of procedures for solutions from more general concepts guided by an appropriate schema. A simple example of the distinction is knowing the rule $7 \times 9 = 63$ as opposed to knowing that one can also get the answer by subtracting 7 from 70. Another is knowing that the area of a triangle equals one-half the base times the altitude instead of knowing that the areas of triangles, parallelograms and trapeziums can all be found by inscribing them in rectangles. The latter type of knowledge is adaptable to wider situations and is more robust in that errors are more obvious and easily corrected.

Most mathematicians would assert that relational mathematics is better than instrumental mathematics because it gives much more information in return for the time invested in learning the relationships. But Skemp and others have observed that a great deal of teaching in elementary and secondary school mathematics is confined to the instrumental level. Teachers report that students feel more secure in learning procedural rules and have an immediate sense of accomplishment when they apply the rule and obtain the correct answer. Relational understandings and larger schema are less well-defined and not guaranteed to lead to the correct solution in every case. At the same time, teachers feel more secure in meeting instrumental goals, where they can measure student progress by simple exercises, than stressing relational understanding, which requires more subtle evaluation.

Nevertheless, the student who will ultimately master mathematics must go beyond these purely algorithmic procedures; it is essential therefore that the assessment of school mathematics be able to gauge the relative numbers of students who are attaining both of these modes of understanding the content. This is the justification of the Computation and Comprehension categories of the SIMS International Grid, and many other assessment schemes. The California Model Curriculum Guide, for example, refers to Number Facts and Arithmetic Operations in instrumental terms and to Mathematical Thinking in relational terms. Similarly, the Illinois assessment labels the instrumental processes as "Computation" and the relational as "Understanding", but it also distinguishes a third category referred to as "Recall", which is defined as a perceptual process of rapid recognition of facts, definitions, and symbols.

Bock & Mislevy (1988) express the instrumental and relational distinction as Procedural Skills versus Knowledge of Facts & Concepts, where facts are understood to mean relational facts, and the terms in which they are expressed, and to exclude purely algorithmic rules. They consider procedural skills to include: calculating, rewriting, constructing, estimating, and executing algorithms. They include in Knowledge of Facts and Concepts, the understanding of terms, definitions, concepts, and principles. We have characterized this category as *Conceptual Understanding.*

The third main category that appears in almost all lists of behavioral objectives for school mathematics is *Problem Solving.* In some lists this category is subsumed under "applications"; in others, under "higher order thinking". It is also a prominent topic in the mathematics education literature of many different countries (Lester & Garofalo, 1982; Schoenfeld, 1985; Kilpatrick & Wirszup, 1969, 1970).

The SMSG and SIMS classify various types of problem solving under both Applications and Analysis. The Illinois and California assessments both use the term "Problem Solving". Bock & Mislevy (1988) include problem solving in Higher Order Thinking, along with Proof, Reasoning, and Real-World Applications.

For the test constructor, the category of problem solving presents difficulties because mathematics educators apply it to at least three different types of tasks. Some assume that it means a mathematician's kind of problem: a proof or a mathematical recreation, a problem that is concise, possibly puzzling, and admits of a clever and elegant solution. Schoenfeld (1985) studies cognitive processes in such problems drawn from plane geometry, while other investigators use simple number-theoretic problems in the same way. There are anecdotal reports suggesting that problems of this type are important in attracting talented young people to careers in mathematics. But there are equally many reports from classroom teachers that other children, talented in their own ways, have no interest in such problems and never gain any skill in solving them.

The second type of problem-solving task is the "story problem"—part of the stock-in-trade of the commercial arithmetic textbook. After procedures have been explained and drilled in the familiar rubrics of arithmetic operations, these textbooks present a story problem requiring their use. The student has to identify the procedure that applies to the situation and insert the numbers in the right places. Many texts teach the "key-word" method of solving these problems: the student is instructed to look for certain words that signal the operation. If the problem asks, "what fraction of a whole pie is one-third of half a pie?", then the word "of" between the fractions indicates that multiplication is required. If the problem asks, "how many eggs are left in a one-dozen box after seven have been used?" the word "left" signals subtraction. The attempt is again to reduce the mathematics to a set procedure. Skemp characterizes it as a very fragile approach that is likely to breakdown as problem complexity increases.

Other educators, including H. O. Pollak (1970) and Max Bell (1972), fault the conventional story problem for its typical triviality. They want to see problems that touch on important real-world applications, have historical antecedents, and relate to a larger domain of ideas. To some extent, this point of view is beginning to change the exercises in texts and workbooks. More realistic and better-motivated examples are appearing in experimental and commercial instructional materials for mathematics; they can, of course, easily be adapted for assessment. Potentially, they could make the assessment tasks more interesting to the students and more relevant to practical works. Many examples of such problems appear in Usiskin and Bell (1983).

An interesting aspect of story problems is that they often involve quantities, directions, measurement and other physical descriptors that are concretely visualizable in space. Many mathematics educators emphasize the facilitating role of spatial visualizing in arithmetic reasoning. Analyzing the role of verbal and visual symbols in mathematics, Skemp (1987) concludes that visual symbols best convey integrative problem structure, while verbal symbols are best for analytical detail. He also observes that, because verbal symbols, in the form of speech, are more efficient for the social exchange of ideas, they have a necessary priority in instruction. But he considers presentations in which both symbol systems interact, as in analytic geometry, to be among the most effective instructional modes.

Bock and Zimowski (1989) present evidence based on the arithmetic items from the California assessment in grade 8 that scores for story problems involving quantity, direction and physical relationships are associated with the overtly spatial items in the assessment. Presumably, the students who benefit from spatial reasoning are those who take a more relational approach to mathematics and do not depend on procedural skills alone. Inasmuch as cognitive studies show spatial ability to be an important source of individual differences, the inclusion of problem solving as a reporting category in the mathematics assessment is further justified on psychological grounds.

Gadanidis (1988) refers to problem solving as the "third dimension of mathematics teaching". His labels for the first and second dimensions are respectively, Facts and Skills, and Understanding. According to his description, the former consists of "routine practice of narrow skills", which would include the procedural skills of Skemp's "instrumental mathematics". The latter corresponds to the relational or conceptual mathematics category identified above. Gadanidis' suggestions for the teaching of problem solving, like those of Schoenfeld (1985), echo Polya's advice on the teaching of heuristic: locate the facts, analyze the problem, keep an open mind, and check frequently one's progress toward the solution. By clever writing of items, the test constructor should be able to tap these processes to some degree. It is relatively easy, for example, to devise geometric problems that have difficult, obvious solutions and easy, subtle solutions. The obvious solutions lead to hidden difficulties, whereas the subtle solutions quickly make the problem transparent. Such items could distinguish the students who perseverate with unproductive ideas from those who realize that progress is slow and that they should try other approaches to the problem solution. See Pandy, 1989, for examples of this more diagnostic style of item writing. Many good examples also appear in the *Taxonomy of Educational Objectives* (Bloom, 1956).

Our final choice of process categories is essentially the same as that of Gadanidis (1988), but we prefer the terms *Procedural Skills, Conceptual Understanding, and Problem Solving.* The first two correspond to Skemp's (1987) categories of Instrumental Understanding and Relational Understanding, where we include accurate knowledge of terms and definitions in Conceptual Understanding. The third category is problem solving, embracing proofs, story problems, real-world problems, and heuristics. With these elaborations, our content-by-process classification of tasks for the eighth-grade mathematics assessment takes the form shown in Table 2.3.

In summary we mean by "Procedural Skills" the ability to perform any of those operations in mathematics that begin from unambiguous givens and lead by a set path to a unique result. This includes the arithmetic operations on given numbers, algebraic manipulation of expressions and equations, constructions with ruler and compass, measuring objects on conventional scales, averaging quantities, *etc.* We deliberately limit this category to the execution of procedures and not the larger understanding of them (which is part of Conceptual Understanding) or the ability to adapt or invent them (which is part of Problem Solving). Our reason for doing so is based partly on factor analytic studies that show fluency of routine symbol processing to be a distinct dimension of individual differences (see Bock & Zimowski, 1989). In addition, we want the scores to be sensitive to instructional emphasis on execution of procedures versus understanding and adaptation of them.

"Conceptual Understanding" means to us comprehension of the origin and significance of the essential results of mathematics. This includes definitions and factual knowledge, not simply as rote recall, but, as conceived by Skemp (1987), as relational knowledge organized in an appropriate schema.

Finally, for "Problem Solving", we have accepted for the Duplex instrument all three classes of tasks that go by that name in math-

TABLE 2.3

Content by classification of tasks for the 8th-grade mathematics assessment instrument

		Proficiencies		
Content Categories		a. Procedural Skills	b. Conceptual Understanding	c. Problem Solving
10.	*Numbers*			
	Integers	11a	11b	11c
	Fractions	12a	12b	12c
	Percent	13a	13b	13c
	Decimals	14a	14b	14c
20.	*Algebra*			
	Expressions	21a	21b	21c
	Equations	22a	22b	22c
	Functions	24a	24b	24c
30.	*Geometry*			
	Figures	31a	31b	31c
	Relations & Transformations	32a	32b	32c
	Coordinates	33a	33b	33c
40.	*Measurement*			
	English & metric units	41a	41b	41c
	Length, area & volume	42a	42b	42c
	Angular measure	43a	43b	43c
50.	*Probability & Statistics*			
	Probability	51a	51b	51c
	Descriptive Statistics	53a	53b	53c

ematics instruction—proofs, story problems, and realistically motivated applications. Although problem solving is not always well represented in multiple-choice achievement tests, there are now many sources of suitable items in the mathematics education literature.

Despite the importance we attach to the difference between procedural Skills and Conceptual Understanding, we appreciate that for purposes of classifying items in the categories in Table 2.3, the distinction is often hard to make. If a task elicits an overlearned operation, so that a successful result is merely a matter of speed and accuracy, then it is measuring a procedural skill. But if the task depends on knowledge of a principle that has not yet become essentially an automatic rule, then conceptual understanding is involved. An example would be knowing that $aaaaa = a^5$: for some students at the eighth-grade level, this would be an absolutely routine algebraic manipulation; for many others, it might require a more searching understanding of the relationship between multiplication and exponentiation. Readers of this report can check our understanding of this distinction by inspecting the items of the instrument, classified by content and process in Appendix B.

Because we are aiming here for an instrument that can be administered in 45 minutes, we do not include enough items to score students on the content subcategories (topics). For purposes of scoring the forms at the student level, we therefore group the cells of the content-by-process array shown in Table 2.4. (At the group level, every cell of the table can be scored.) The score profiles for the student reports thus consist of the five main content categories (Numbers, Algebra, Geometry, Measurement, and Probability and Statistics) and the three process proficiencies (Procedural Skills, Conceptual Understanding, and Problem Solving). For reporting purposes, we will shorten the term Probability and Statistics to "Statistics", Procedural Skills to "Procedures", and Conceptual Understanding to "Concepts". Examples of reporting forms containing these labels appear in Chapter 7.

Like other investigators, we have chosen the content and process categories on formal and theoretical grounds, and because they are widely recognized by mathematics education. But it remains an open question whether these distinctions correspond to empirically demon-

TABLE 2.4

The content by process array for grade 8 mathematics field trials

		Proficiencies		
Content Categories		a. Procedural Skills	b. Conceptual Understanding	c. Problem Solving
10.	Numbers			
	Integers	11a	11b	11c
	Fractions	12a	12b	12c
	Percent	13a	13b	13c
	Decimals	14a	14b	14c
20.	Algebra			
	Expressions	21a	21b	21c
	Equations	22a	22b	22c
	Functions	23a	23b	23c
30.	Geometry			
	Figures	31a	31b	31c
	Relations & Transformations	32a	32b	32c
	Coordinates	33a	33b	33c
40.	Measurement			
	English & metric units	41a	41b	41c
	Length, area & volume	42a	42b	42c
	Angular measure	43a	43b	43c
50.	Probability & Statistics			
	Probability	51a	51b	51c
	Descriptive statistics	52a	52b	52c

strable dimensions of mathematics performance. Conceivably, contrasts among them make no useful distinctions between students, or between schools or programs. Obvious examples lead us to think otherwise, however. Certainly, profile differences with respect to content should appear between schools that do or do not include algebra or pre-algebra among their mathematics classes. Similarly, we might expect geometry and statistics to be depressed in states where neither are consistent parts of eighth-grade math curricula or instruction.

As for process differences, studies in the literature have contrasted traditional tests that emphasize procedural skills with those that tap conceptual understanding. In one of the few randomized studies of mathematics-instruction programs, Milton Maier found that relative performance on these types of tests strongly discriminated between traditional instruction and that based on the School Math Study Group materials (see Bock, 1975, p. 236). Similar evidence of effects involving problem solving versus the other process proficiencies should become apparent as the movement for greater emphasis on realistically motivated exercises and applications begins to influence teaching practices.

We also report in Chapter 7 some results from California schools showing that classrooms taught by different teachers have discernibly different content and process profiles, possibly reflecting different teaching styles or emphases. Findings of this kind reinforce a conception of school mathematics attainment as multidimensional—something that cannot be described as a single score, but only as a set of scores reflecting the lines along which the subject divides in instruction and in learning. The Duplex Design is structured to provide this multidimensional account in less and greater detail in the student-level and school-level reports, respectively.

2.2 Principles of scoring the Duplex Design

To provide diagnostic scoring at more than one level, we have applied to the Duplex Design a scoring scheme based on a hierarchical model for variation in the observed responses. Except for the state mean, each term in the model is considered an independent source of variation expressed as a deviation from the term at the next higher level. Listed from the highest to lowest levels, the terms in the model for a student response in a particular cell of the content-by-process classification are as follows:

1. The state overall mean-score for mathematics
2. The state profile-score for the content-by-process combination
3. The school overall mean-score for mathematics
4. The school profile-score for the content-by-process combination
5. The student's overall mathematics score
6. The student's profile-score for the main content category
7. The student's profile-score for the process category
8. The student-by-content-by-process interaction

Level 8 of this hierarchy is the source of error variation in the student-level scores. The corresponding error standard deviation is estimated along with the profile scores and enters into the calculation of the standard errors of measurement for the scores. The size of these standard errors depends upon the number of student-by-content-by-process interaction terms that are aggregated in computing the student's scores. In the design shown in Table 2.3, there are more content-by-process subcategories within the main process categories than within the main content categories. Consequently, the measurement errors for the process scores will be smaller than those for content. This will be apparent in Chapter 5 when we examine the student-level scores from the California field study.

At the school level (Levels 3 and 4), we estimate scores for each of the content-by-process combinations that define the cells of the Duplex Design. As discussed in Chapter 1, these estimates aggregate responses across the multiple forms that make up the assessment instrument. At this level, all lower-level sources involving the students are sources of sampling variation. Thus, the standard errors of measurement of the school scores depend upon the number of students tested in the school as well as on the number of forms. The number of forms determines the number of independent student-by-content-by-process interaction terms that are aggregated in the school-level scores, which in turn affects the generalizability of the scores.

In respect to census assessments at the selected grade-level, it may seem strange that we considered the students a sample for purposes of computing an error estimate for the school-level scores. But the statewide annual cohort is indeed a sample of the population of successive cohorts of students from the school's catchment area, and this is the relevant population for purposes of defining trends. The scores that are estimated for the school are intended to characterize typical performance, not just in the current year, but as a trend extending over a number of years. With respect to this longer-range tendency, the students in any particular year are merely a sample from the larger population over time. Any given student can reasonably be regarded as independently sampled from that population if minor sources of correlation such as the presence of siblings in the school are ignored.

The effect of school size on the accuracy of estimating the school-level scores can be a problem for assessment in states, such as California, that have numerous schools with very small numbers of students per grade. In some cases the scores for the detailed content-by-process subcategories for these schools are not determined accurately enough to warrant reporting. In those cases, the school-level reports have to be limited to main categories and overall scores. Fortunately, student-level reports are not affected by small school size.

At the state level, all lower levels of the scoring hierarchy are considered sources of sampling variation. But in a census assessment, the numbers of students is typically so large that sampling variability from that source can be virtually ignored. It remains possible, however, that in a small state the variability of school effects over years could be a source of instability in the estimated state mean-scores. At present, the types of multilevel analysis required to estimate the relevant variance components are not available, but recent progress in multilevel analysis of educational data should soon show whether variation from these sources is appreciable (see Bock, 1989).

As we discuss in Chapter 3, the scales in terms of which we will report the assessment results have neither a natural origin nor unit of measurement. They must be assigned arbitrarily in the first year of of the testing indicated that the opportunity to explain the purpose and nature of the test on the first day, and to allow the students to try a sample of items in the pretest, made the second-stage testing less stressful for the students than is typical of other external testing. If the test had been given in one stage, the forms would have to have been much longer in order to include items suitable for all levels of ability. From the point of view of both the teachers and the students, working with two relatively short test forms on two different days was much to be preferred to a longer test, comparable to a traditional achievement test, on one day.

In the design for eighth-grade mathematics studied here, the second-stage test consisted of 24 test booklets organized into eight forms of three booklets each, one pitched at an *easy* level of difficulty, one at a *medium* level, and one at a *difficult* level. The two-stage structure is represented schematically in Figure 2.1. The main advantage of two-stage testing is the almost total elimination of floor and ceiling effects (students scoring all incorrect or all correct). These effects, especially troublesome in short scales such as those in the Duplex instrument, can lead to U-shaped distributions of student-level scores, making data analysis very difficult. Even a two-stage test with only three second-stage forms is a marked improvement over a one-stage test in this respect (see Lord, 1980).

Toward the center of the score distribution, the gains in efficiency (as measured by the ratio of the error variance of the two-stage test to that of the one-stage test) is smaller, but it still favors the two-stage procedure. We present an analysis of the efficiencies of the California version of the two-stage Duplex instrument as a function of the student-level scores in Chapter 5.

It has been suggested that teacher judgments could be substituted for the first-stage test; that is, the teacher would assign the second-stage test booklet on the basis of the student's previous grades and classroom performance. Because it would introduce an element of subjectivity in the scoring procedure, however, we did not consider

procedure resulted in useful gains in efficiency over one-stage testing, but it proved too complicated for the sixth-grade students on whom it was tested. Because it also had the disadvantage of added printing costs due to including all the items in every second-stage test booklet, it was not pursued further.

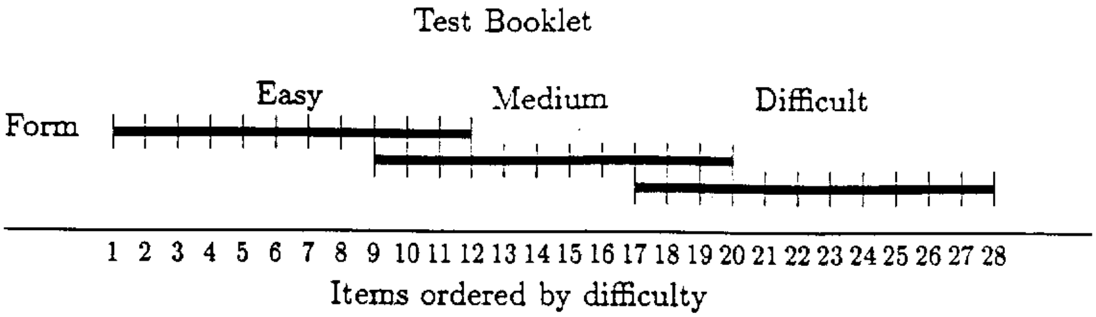


Figure 2.1. Schematic representation of item assignment to the second-stage test booklets for one of the scales of the Duplex instrument. The bars cover the items included in the three test booklets of one of the eight forms of the instrument.

the use of teacher judgments advisable in the heterogeneous population of schools found in Illinois and California. But if objective test scores relevant to the subject-matter were available for the students, there would be no objection to equating them to the first-stage test scores for purposes of making the second-stage assignments without administering the first-stage test.

2.4 Summary

The main measurement innovation in the Duplex Design is the provision for separately evaluating the contributions of content knowledge and process skills to individual student performance of the assessment tasks. The evaluation depends upon a cross-classification of each item in the Duplex instrument into a content and process category. In the context of eighth-grade mathematics, mathematics educators are broadly in agreement that the relevant content categories are 1) Numbers, 2) Algebra, 3) Geometry, 4) Measurement, and 5) Probability and Statistics. Although terminology differs, there is comparable agreement that the main process categories for school mathematics should be 1) Procedural Skills, 2) Conceptual Understanding, and 3) Problem Solving. In the Duplex Design, the content categories are subdivided into *topics*, each represented by suitable items, but the process categories are not further divided. In the design prototype of the present study, items representing all combinations of the content topics and the processes appear. It is not essential in a Duplex Design that items exist for all such combinations, some may be logically impossible, but the analysis is simplified if the cross-classification is essentially complete.

Scoring of the prototype instrument occurs at two levels—the student level and the school level. At the student level, the scores measure performance in the five main-content categories and in the three process categories. Content scores aggregate over processes, and process scores aggregate over content. Aggregation over both content and process gives an overall mathematics score for the student.

At the school level, attainment in classrooms and for the school as a whole is measured in each of the content topic-by-process subcategories. These scores measure attainment of detailed curricular objectives, including effects of the interaction of content and process, and can also be combined into scores for the district or higher level of score aggregation, including the state.

To provide reliable scores at the student level, the Duplex Design requires more student time for testing than a pure matrix-sampling design reporting only at the school level and higher. The amount of time per student can be reduced, however, through the use of two-stage testing. Students take a short pretest that routes them to a second-stage test adapted to their general levels of proficiency in the subject-matter. The second-stage form can be shorter, and still be reliable, because nearly every item is informative for the particular student. Because the pretest provides an introduction to the test content, and the second-stage items are better suited to the student's capacities, students are better motivated and more accepting of a two-stage test than a one-stage test. Two-stage testing was a feature in both the field trials of the present study; an evaluation of the operating characteristics of the procedure appears in Chapter 5.

Chapter 3

Analyzing the Duplex Design

Considering the large number of items that comprise an assessment instrument, it is not practical to think of reporting results for individual items except in the form of detailed item statistics intended for specialists in test construction and psychometrics. All other uses of the data necessarily depend upon summary statistics, which we generally call "scores", that express the item information in more manageable form.

In the field of educational measurement there are at present two quite different approaches to computing such scores. One is called "domain-sampling theory", and the other, "item response theory", or "IRT". Both have been used in reporting assessment data. In this chapter, we explain the essential differences between these two theories and give reasons why we have adopted IRT methods to analyze data from the Duplex Design.

3.1 Domain-sampling theory versus IRT theory

In ***domain-sampling*** theory, any given educational objective is assumed to specify a domain of tasks or items by which mastery of the objective can be assessed. A particular test of the objective is assumed to be a sample from the population of all possible items composing the domain. This purely conceptual population is referred to as the "item universe". If the items of the test are considered a probability sample of this universe, the student's percent-correct score on the test is a best estimate of the percentage of items in the universe to which the student could respond correctly. Thus, to the extent that the item content of the domain is well-defined, the quantity estimated by the student's percent-correct score has a clear meaning.

Domain-sampling theory does not necessarily require the content of the domain to be homogeneous; the scope of the domain is purely a matter of definition. Thus, the entire content of eighth-grade mathematics specified in Table 2.2 might be assumed the domain for estimating the universe percent-correct score in the subject matter. Concurrently, the marginal content or process categories might define the domains for a report of a student's strengths and weaknesses in areas within the subject matter. Even the individual cells of the content-by-process classification could be domains for purposes of a detailed evaluation of instruction. Domain-sampling concepts apply at any of these levels.

A simple example of a well-defined domain in another area is a list of words that a person should be able to spell correctly. The word books compiled for secretaries are examples of such lists. Because the universe is exhaustively defined, a spelling test consisting of a random sample of words from such a book has a clearly interpretable domain percent-correct score. If the book is assumed to contain all the words that a secretary will have to transcribe from dictation, a score of 98 percent on the test is an estimate that, out of every 100 *different* words dictated, the secretary is expected to misspell 2. Thus, a score of 98 percent on the test might be a reasonable requirement for employment.

The domains representing most educational tests are not as clearly defined, but in principle they can be understood as a large set of objectively defined tasks, possibly structured in some logical system (items are often hierarchically organized.) To the extent that two independent test constructors are working from the same domain definition, they should be able to construct tests that estimate the same domain percent-correct score, just as two independent sample-survey experts should be able to reach the same estimate of, for example, the number of persons older than age 30 who have never married. This sampling conception of an estimated percent-correct score is represented schematically in Figure 3.1.

In *item response* theory, in addition to an assumed domain of

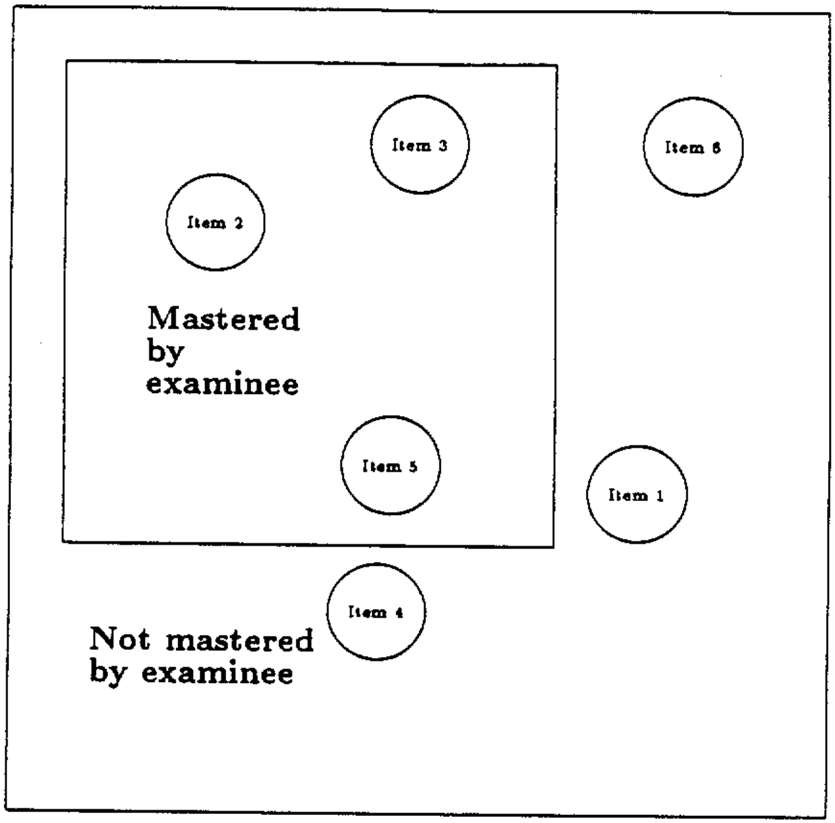


Figure 3.1. Representation of a domain percent-correct score. The score is the area of the smaller box relative to that of the larger.

items, a population of respondents (in the present context, the students) is also assumed. The items in the domain are assumed to be structured, not only by logical categories, but also by their location along a linear continuum related to their average difficulty for the respondents in the population. Statistical procedures are available to estimate the precise location of any item on the continuum from the responses of a large sample of respondents. These locations are very nearly in the same order as the so-called "item p-values", *i.e.*, the percent of correct responses to the item as observed in the sample of respondents. They are one of several characteristics of items (others are discriminating power and probability of chance correct responses) referred to as "item parameters". The procedure for estimating them is called "item calibration". A review of the essentials of item response theory may be found in Lord (1980).

The concept of an examinee's score is quite different in item response theory than in domain-sampling theory. In IRT, a score is expressed by the location on the continuum from easy to difficult items where the examinee's probability of answering correctly the items is near 50 percent. The point that locates the examinee on the continuum is referred to as an "ability" or "proficiency". The estimate of this point is called a "scale score". This conception of a scale score is illustrated in Figure 3.2.

If the test items are a representative (probability) sample from a specified universe as assumed in domain-sampling theory, the item parameters will have a distribution in the domain universe. From this distribution, item response theory makes it possible to calculate from the scale score the percentage of items in the domain that the examinee could be expected to answer correctly. Thus, item response theory can include the concept of the domain percent-correct score.

When used in this way to serve the purpose of domain-sampling theory, IRT requires only one additional assumption beyond those of domain-sampling theory. It assumes a statistical model that accurately predicts the probability of a correct response as a function of the item parameters and the examinee's position on the ability continuum. These so-called "item response models" (IRM's), or "item response theoretic" (IRT) models, constitute the central and distinctive feature of item response theory. Models are available for most common types

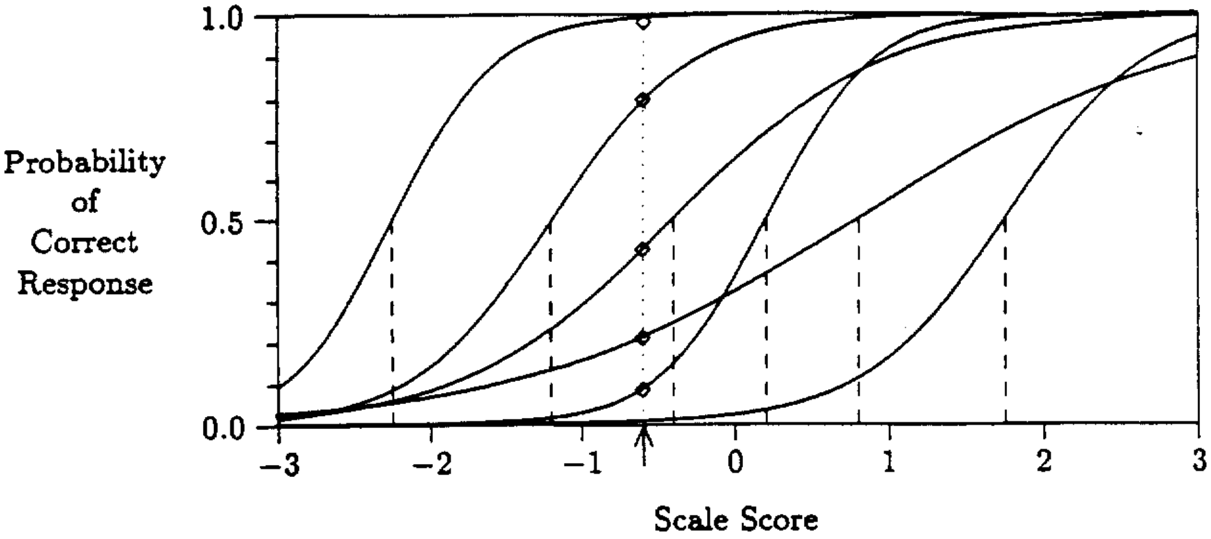


Figure 3.2. Representation of a scale score. The score is the location of the examinee with respect to a linear ordering of the domain content. The probability that the examinee has mastered each item in the domain is the height of each item response function at the examinee's scale-score location. The sum of the probabilities gives the examinee's domain percent-correct score.

of test items, including short-answer and multiple-choice items scored right, wrong, or omitted; graded items used in ratings of essays and other performance tasks; and items with multiple nominal-response categories (see Thissen & Steinberg, 1986). Well-developed computer procedures now exist for item analysis and test scoring based on these models. The models we employ in the present study are described in Chapters 5 and 6.

One might well ask why IRT methods should be used to estimate a domain percent-correct score when the sampling approach is so much simpler. The reason is that it is easier to obtain comparable measures when alternative forms of a test are scored by IRT methods rather than number-right scoring. In principle, comparable scores can be obtained under domain-sampling assumptions simply by randomly sampling items from the universe for each of the forms.

Unless the forms have many items, however, there will be variability in the difficulty of the forms due to item-sampling variation. Again in principle, this can be avoided by very carefully stratifying the items with respect to their population percent-correct values (p-values) and reliability indices. But in practice, this stratification is very difficult; there are seldom enough items in an item pool to obtain enough good matches, especially in the Duplex application where we require a large number of forms. Situations can also arise in which between-form variability is increased because new items are added to the pool after some of the forms have been prepared. The classical method of handling these problems is to allow the difficulties of the forms to vary but to construct tables for converting the scores to common values by so-called "equipercentile" equating.

But the IRT approach is even simpler: the forms are constructed by sampling from the item pool (with stratification by content and process in the case of the Duplex Design); then the sampling variation between forms is accounted for by the scaling procedure that adjusts for the empirically determined characteristics of the items. There is no necessity of matching items during test construction or of constructing equating tables. The forms are simply administered randomly to members of the respondent population (in school settings this is done by assigning the forms to the students in rotation), and the item characteristics are estimated from the responses. The IRT scoring procedure then makes use of the estimated item-characteristics to compute scores on the same scale from all forms.

The logic of this IRT procedure is analogous to that of survey sampling when statistics from an allocation sample are converted to those for a probability sample by use of case weights based on demographic characteristics of each respondent. Although these statistics could also be obtained by matching cases to population values for the demographic categories, the case-weighting method, like IRT scoring, makes use of all the data and is a simpler procedure.

Another advantage of IRT scoring under domain-sampling assumptions is that the scale scores for two different tests are more likely to be linearly related, than are percent-correct scores. The reason is that the particular distribution of item difficulties in each test distorts the underlying bivariate relationships between the attainment variable in a complex way. These "typical distortions" of cognitive measurement have been extensively investigated by Frederic Lord (see Lord and Novick, 1968). Because they are largely eliminated from the relationships between IRT scale-scores, the statistical analysis and graphical representation of test or assessment results are simplified and easier to interpret.

Moreover, the IRT treatment of the data allows us to take advantage of adaptive testing, such as two-stage testing, to improve the efficiency of assessments. Results of the two-stage testing can be expressed on a common scale regardless of whether the student has been assigned the Easy, Medium, or Difficult second-stage test booklet. We discuss the scoring of the two-stage Duplex instrument by this method in Section 3.4. Finally, only IRT permits the items and the respondents to be located on the same scale for purposes of a content reference interpretation. In Chapter 7, we make use of this felicitous property of IRT scale-scores in our proposals for reporting Duplex results.

3.2 Student-level scale scores

For the student-level content and process scores of forms constructed according to the design in Table 2.4, strict unidimensionality of the item responses probably can not be assumed. In each content area, the items belonging to the same proficiency probably would be somewhat more highly associated than those belonging to different proficiencies; similarly, in the proficiencies, items belonging to the same content would be somewhat more highly associated. (Those effects can be examined empirically by item factor analysis; see Bock, Gibbons, & Muraki 1988). This is the reason for maintaining exactly the same content and process structure in all the forms. It provides the stratified sampling design that permits the content and process domains to be defined by domain-sampling theory. IRT methods of scoring merely make the scores comparable from form to form.

If standard IRT procedures are employed in calculating the student-level scores, however, the standard errors of the scores will be underestimated if there is greater association among item responses within the content or process categories than between categories. Thissen, Steinberg & Mooney (1988) have shown that this problem can be avoided by treating the items within each category as a "testlet", as defined by Wainer & Kiely (1987). For example, the procedural items within Numbers would constitute one testlet, those within Algebra another, *etc.* The IRT analysis would then assume unidimensionality *between* testlets, but not between items within testlets. The IRT model for multiple categorical responses would directly express the probability of each pattern of correct and incorrect item responses within the testlet. The student-level scores for Procedures, or any of the content or process categories, computed in this way would then have the correct standard errors (see Thissen, Steinberg & Mooney, 1988).

The IRT model for multiple nominal categorical-responses introduced by Bock (1972) is suitable for this type of testlet. Each possible pattern of correct and incorrect responses within the testlet is treated as a nominal category; then parameters of a linear model are estimated so as to best account for the observed frequencies of patterns in the data from a large sample of respondents. The marginal maximum likelihood method of Bock & Aitkin (1981) is used for this purpose.

At present, the computer programming for the testlet method of scoring is not available. We have therefore used conventional IRT methods both in fitting the response models for the items and in computing scores for students. As a result, the standard errors for these scores may be somewhat underestimated, but the results of studies by Gibbons, Bock & Hedeker (1988) suggest that the effect of these failures of unidimensionality is rather small.

3.3 School-level scale scores

In school-level scoring, the domains defined by each of the content-topic and process cells of the Duplex Design are sufficiently narrow to justify the assumption that the corresponding item universe is unidimensional. That the assumptions of IRT can also hold in group-level assessment was first pointed out by Bock, Mislevy & Woodson (1982) and investigated in detail by Mislevy (1983). A necessary condition for group-level application of IRT is that each item of any given scale appears on a different form of the instrument. In this way, each item response comes from a different student, and the responses are experimentally independent. The data are then amenable to standard IRT estimation methods that assume independence.

The chief difference is that the data for the school-level analysis take the form of the numbers of students attempting and responding correctly to each item, while the data at the student-level are the binary item scores (1 if correct, 0 if incorrect). Technically, and under the corresponding independence assumptions, the school-level data are *binomial* variables and the student-level data are *Bernoulli* variables. IRT models are easily formulated for either of these types of data (see Mislevy, 1983).

In student-level data, the IRT model describes variation underlying the item response that arises out of the interaction between a specific respondent and a specific item at a specific time. This variation can reasonably be assumed to be the additive result of many finite, more or less independent, influences and, thus, to be normally distributed. This is the justification of the cumulative normal distribution function (or its almost identical proxy, the logistic function) as the IRT model for individual-level data. To use the same models at the group-level, a further assumption is required—namely, that the proficiency measured on the underlying scale is normally distributed within the group. This is also a reasonable assumption if the group is more or less homogeneous, which is generally the case within classrooms or schools. The assumption should, of course, be tested empirically in the course of fitting the IRT model. We have carried out such tests extensively in data from the California assessment and found only scattered instances of significantly poor fit. Of the few cases of model failure that have been detected in very large samples and among many items, almost all could be attributed to ambiguous alternatives of multiple-choice items (Schilling & Bock, 1989). Rewriting of such items will generally lead to acceptable fit. In the data of the present study, only a few instances of marginal fit were found, but the sample size per item was not large enough to permit a sensitive test of the group-level model (see Chapter 6). If a three-parameter logistic model is assumed for the group-level analysis, as in this study, the marginal maximum likelihood method of estimating item parameters (in the population of schools) and the Bayes methods of estimating school scale-scores are identical to those of the more familiar student-level analysis, except that a binomial frequency function is substituted for the Bernoulli frequency function of that analysis. The BILOG computer programs gives the user the option of this substitution.

As in student-level IRT analysis, the origin and unit of the scales for the school-level scores are arbitrary. In the California assessment, which does not include student-level scores, origin and unit are set so that the distribution of school-scores, weighted by the number of students in the school, have mean 250 and standard deviation 50. Rather than adopt this convention, we have chosen to allow the scales of the respective student-level scores to determine the scales of the school-level scores. Mislevy & Bock (1989) have proposed an IRT model integrating both of these levels of scoring into a single analysis on a common scale, but it is not yet implemented for practical use. We have chosen instead to determine a *between-school* interaction component of variance for the student-level scores and set the unit of the school-level scores so that their between-school interaction component is the same value. The origin of the school-level scores is then set so that the estimated weighted mean of the *school-level* scores is equal to the state mean of the *student-level* overall mathematics score (details of these conventions are presented in Chapter 6.)

3.4 Scoring two-stage tests

Adaptive testing, including two-stage testing, has no counterpart in classical test theory. It is possible only because IRT scale-scores of the respondents can be computed from arbitrary subsets of items from the test. These computations require that the parameters of the items be estimated with respect to the full test, but this is readily accomplished if the item subsets that make up the second-stage tests are connected by common "link" items (see Figure 2.2).

When the MML method is used, the procedure for estimating the item parameters of the second-stage test is essentially as follows. Initially, the parameters are estimated separately in each of the groups of respondents who were administered a distinct second-stage test booklet. In the present study, these groups corresponded to the Easy, Medium, and Difficult booklets. Because the proficiency distributions for these groups are arbitrary and unknown, each is estimated along with the item parameters of the respective group. But since the origin and unit of scale of the separate analyses is indeterminate, a basis must be found for adjusting the separate sets of parameter estimates so that they are expressed on a scale with common origin and unit.

This is where the link items come in. We first set the scale of the three distributions to the same value by constraining the slopes of the link items in the corresponding MML analyses to be equal. At the same time, we also constrain the lower asymptotes, or "guessing" parameters of these items to be equal. Under these restrictions, the average of the differences between the estimated location parameters, or "thresholds", of the link items that are common to two of the groups provides an estimator of the difference between the corresponding distributions. When the locations of all the items, both link and non-link, are adjusted by an amount equal to half this difference, their response functions are all brought onto the same scale. They are then in a form suitable for computing comparable scores for the respondent, regardless which second-stage test they were administered. We give the details of this procedure, modified somewhat to allow for specific features of the pretest applications, in Chapter 5. A similar procedure for estimating parameters of the school-level IRT model is discussed in Chapter 6.

Once the items are calibrated in this way, the calculation of scale scores for the respondents is a straight forward application of IRT scale-score estimation. As we explain in Chapter 5, we use Bayes estimation procedures for this purpose because they have better properties than other methods when the scores are based on relatively small numbers of items, as is the case in the Duplex Design. Similar methods apply to the estimation of school-level scores for curricular objectives (see Chapter 6).

3.5 Modeling item-parameter drift

The IRT scoring of assessment data normally makes use of an item calibration in the first fully operational year of the assessment instrument. Thereafter, the scores refer to the scale definition established in the base year. Any interpretation of the scale scores in subsequent years is relative to the status of the population in that year. In this respect, the IRT calibration is similar to establishing percentile norms for test scores in a given year. Like test norms, however, IRT calibrations can become out of date over extended periods of time. For the test as a whole to change in difficulty is not a problem; these effects are absorbed in the mean score for the population. Such overall tendencies are in fact the changes that the assessment is designed to monitor..

But it may also happen that some of the item parameters, especially those reflecting item difficulty, change differentially as emphasis on different subject matter topics change. This phenomenon is referred to as "item-parameter drift" (Goldstein, 1983). That such effects actually exist, and can be described and modeled, has been demonstrated in data from the College Board Physics Achievement test by Bock, Muraki & Pfiffenberger (1988). These authors examined data from a form of this test that Educational Testing Service administered to large national samples on five occasions between 1973 and 1982. Working with the longest homogeneous topic within the test (Mechanics), they fitted and tested a time-trend IRT model to items from this content. Their analysis showed appreciable differential drift in the location parameters of these items, but no significant drift in the item validities (discriminating powers) or the random success probabilities (guessing parameters). Changes in the location parameters were essentially linear, and the relative directions of change were interpretable in terms of trends in the teaching of secondary school physics in the United States over this period. The authors also suggested how their time-dependent IRT model implemented by the BIMAIN computer program of Muraki, Mislevy & Bock (1987) could be used to monitor differential drift in item locations and to adjust from year to year for its influence on the scale scores so affected.

In the context of the Duplex Design, item-parameter drift is most likely to occur in student-level scores where the item content is relatively heterogeneous. It would not be expected to any extent in the school-level scores for the extremely homogeneous content-topic-by-process subcategories. Differential change of instructional emphasis within these narrow subcategories is unlikely because, while the topic as a whole might receive relatively more or less attention, individual items would not. Bock, Mislevy & Woodson (1982) have referred to such subcategories as "indivisible curricular elements". Computing scale scores for the schools from these homogeneous elements is a source of great robustness and stability in IRT scaling of school-level assessments. No evidence of differential item-parameter drift has been found in the California assessment instruments, which are scaled in similar narrow elements (Mislevy & Bock, 1983).

3.6 Summary

Large-scale educational assessment as presently practiced relies heavily on the responses of students to brief tasks that can be quickly scored. Ratings of more extensive productions of the students are confined almost entirely to written essays in direct writing assessments. Measurement of curricular objectives by means of multiple-item tests is based on the assumption that the subsets of items within the test are representative samples from specified content domains. The student's responses to the items provide an estimate of the percentage of the domain content that the student can be considered to have mastered. We argue here that the estimation of such percentages is better accomplished by item response theoretic (IRT) methods than by the item percent-correct score of classical test theory. Our reasons are that IRT procedures provide: 1) more stable estimation of the domain scores through the intermediate calculation of scale scores, 2) easier student-level equating of the multiple forms that make up the Duplex instrument, 3) easier updating of the instrument without loss of temporal continuity of the assessment scales, 4) the availability of two-stage or other adaptive testing, and 5) locating of items and students on the same scales for purposes of content-referenced interpretation of the scores.

Straightforward extension of the response models to group-level data make IRT scoring at the school level available for the matrix-sampling component of the Duplex Design. These models assume one item per form for each of the scales reported at the school or higher level. This assumption is built into the Duplex Design and allows the reporting of progress in as many curricular objectives as the forms have items, typically, 20 to 50. A number of new developments in IRT are promising for assessment applications. One of these is the modeling of responses to small subsets of related items called "testlets". If the items for topics within the main content categories were considered a testlet, the assumption of overall unidimensionality for the student-level scoring could be relaxed. More accurate estimates of standard errors of the student scores would result (see Wainer & Kiely, 1987; Thissen, Steinberg, & Mooney, 1987).

Finally, new procedures for detecting and accounting for differential item-parameter drift aid in maintaining assessment instruments over extended periods of time while retaining full comparability of scores.

Chapter 4

Overview of the Illinois and California field trials

The feasibility study of the Duplex Design in eighth-grade mathematics was based on two field trials, one in Illinois in the autumn of 1986 and the other a year later in California. We took advantage of the year between the trials to revise the instrument on the basis of the item analysis of the Illinois data, and we improved the training and reporting procedures for the field work. In this chapter, we present an overview of the field trials, focusing on the California version of the instrument, which was the source of the final data reported here.

4.1 The sample design

Samples of 32 schools were selected in each of the states by the NORC field-study director, who worked from school rosters and background data supplied by the state assessment programs. A stratified random sampling procedure was used to select the sample of schools in the following way. All public schools in the state with eighth-grade classes were classified according to a 2^5 design of demographic and socioeconomic background variables. In Illinois the background variables were 1) enrollment size, 2) northern or southern location in the state, 3) percent minority enrollment, 4) urban or rural location, and 5) median income of the school district. In California, median income was replaced with a more general index of socioeconomic status (SES). In each state, one school was randomly selected from each nonempty cell of the respective design. In Illinois, the design yielded eight empty cells; in California, four empty cells. In both cases, schools were randomly selected from adjacent cells to complete the samples.

4.2 Creating the Duplex Design instrument for eighth-grade mathematics

In constructing the assessment instrument for the trials, we conformed to the requirements of the Duplex Design discussed in Chapter 2. The design relies on a multiple-matrix sampling procedure to assess with good generalizability the detailed curricular objectives at the school or higher level of aggregation, while employing two-stage testing to provide reliable measurement at the student level. Implementation of the former requires many distinct forms of the test instrument; that of the latter requires a pretest that routes a pupil to a second-stage booklet matched to his or her level of ability, and within the second-stage forms, to test booklets that vary in their level of difficulty.

Within the constraints of the Duplex Design, we generated the assessment instrument from the content-by-process specification described in Table 2.3 by the following steps. Math items were pooled from the Illinois 8-th grade and 11-th grade assessments, and from the California 8-th and 6-th grade assessments. In addition, Kenneth Travers, of the University of Illinois, Urbana, kindly supplied the items of the Second International Mathematics Study. All of these items had either p-values or IRT item-statistics for the corresponding populations. A committee consisting of John Dossey, Tej Pandey, and Darrell Bock then assigned each item to a content topic and process category. In cases of disagreement, majority ruled or the classification was revised. The goal was to obtain sufficient items for 24 replicate booklets (eight distinct forms, each with three booklets at different levels of difficulty). Although the pool of items was large, there were very few items representing *Irrationals, Inequalities, Other Systems of Measurement*, and *Experiments and Surveys*. As a result, these topics were omitted from the instrument. To enlarge the item pools of other curricular objectives with too few items, John Dossey and his colleagues in Mathematics Education at the University of Illinois, Normal, wrote new items using the items from the other sources as guides.

On the basis of the item-statistics from the Illinois, California, and IEA programs, and by judgment in the case of the new items, the item pools for each curricular objective were stratified into three levels of difficulty—easy, medium, and difficult. At each level of difficulty, one item was selected from the item pools representing each of the 45 cells of the design to produce the easy, medium, and difficult booklets of each form. This process was repeated until eight replicate forms of the test were produced.

Within each form, links among the booklets were established by replacing two items within each proficiency with two items from the booklet or booklets at adjacent levels of difficulty. For example, two items in the Procedural Skills subtest of the Easy booklet were replaced with two items from the same subtest in the Medium booklet and vice-versa. The common items among booklets were selected in such a way that they also provided links among the five content areas.

Finally, twelve items were drawn from the remaining items for the pretest. The items were selected for uniform spacing and high validity and spanned a wide range of difficulty.

This version of the instrument was field-tested in Illinois. Although the pretest functioned largely as anticipated, assigning students to the second-stage booklets in the desired equal proportions, the presence of one item that was too difficult for the students and two with relatively poor discriminating ability, indicated that it could be improved. These poorly performing items were replaced and the test was lengthened to 15 items to enhance the assignment of pupils to the second-stage booklets. The cutting points for this assignment were adjusted to reflect the altered content and length of the pretest. The Illinois study also revealed that the provision for teachers scoring the pretest could be improved. Several teachers reported that the opaque template used to score the pretests was difficult to align with the answer sheets. To minimize this problem, we prepared transparent templates for the California study (Appendix C).

The Illinois study also revealed that the difficulties of some of the items in the second-stage forms were not in accord with their classification as Easy, Medium, or Difficult items. To improve the assignment of items to booklets, the items in the second-stage forms were rearranged for the California study on the basis of the item-parameter estimates obtained in Illinois. The wording and appearance of some of the items were also improved and additional link items were added to the forms.

The item structure of the revised version of the instrument is shown in Table 4.1. The numbers in the table refer to the placement of the items in the booklets of each form. The nineteenth item in each booklet of Form 1, for example, belongs to the Procedural Skills by Algebra-functions cell of the process-by-content design. The link items are labeled EM and MD. EM indicates that the item is the same in the Easy and Medium booklets of the form, MD that the item is the same in the Medium and Difficult booklets. The table also shows that the items in the forms are grouped by content area and rotate through the process categories of each topic. The forms all begin with Numbers, but differ in their ordering of the other content areas.

In preparation for the California study, NORC printed copies of the revised versions of the pretest (Appendix A) and the second-stage booklets (Appendix B) from camera-ready boards. The answer sheet (Appendix D) for the two-stage booklets was redesigned to accommodate the lengthened pretest and copies were printed by National Computer Systems. A guide for the test administrators (Appendix E), a teacher feedback form (Appendix F), and school and class transmittal forms (Appendix G) were also prepared and reproduced at NORC in Chicago. The guide explained the purpose of the study and the procedures for administering the test. The feedback form asked the teachers to comment on the test and the testing procedures. The class transmittal forms asked the teachers to identify special students—honors/advanced placement, special education, learning disabled, and English as a second language (ESL)—who were members of their class.

4.3 Field procedures in California

In the California study, NORC field personnel made their initial contact with the schools in early September of 1987 and arranged for testing dates and in-person visits to brief the school personnel on the

TABLE 4.1
Item Structure of the eight forms

	Form							
Curricular Objective	1	2	3	4	5	6	7	8
Procedural Skills								
Numbers								
Integers	1MD	1MD	1	1	1MD	1EM	1MD	1MD
Fractions	4MD	4	4MD	4MD	4	4	4EM	4
Percents	7EM	7EM	7	7EM	7EM	7MD	7	7
Decimals	10EM	10MD	10MD	10MD	10	10	10	10
Algebra								
Expressions	13	13MD	37EM	37MD	28	28	19	19EM
Equations	16MD	16	40	40EM	31MD	31MD	22	22EM
Functions	19	19EM	43EM	43	34MD	34EM	25EM	25
Geometry								
Figures	22	22MD	13EM	13	37MD	37MD	28	28
Relations & Transformations	25	25	16	16	40MD	40EM	31MD	31MD
Coordinates	28	28	19	19	43EM	43EM	34EM	34MD
Measurement								
English & Metric Units	31EM	31EM	22	22	13	13MD	37EM	37
Length, Area & Volume	34EM	34EM	25EM	25	16	16	40	40MD
Angular Measure	37MD	37	28MD	28EM	19	19EM	43	43
Probability & Statistics								
Probability	40	40	31EM	31EM	22EM	22MD	13MD	13MD
Descriptive Statistics	43	43MD	34MD	34EM	25EM	25	16MD	16EM
Conceptual Understanding								
Numbers								
Integers	2MD	2MD	2MD	2	2EM	2	2	2
Fractions	5	5EM	5MD	5EM	5MD	5	5	5
Percents	8MD	8	8	8	8	8EM	8EM	8EM
Decimals	11EM	11EM	11EM	11EM	11	11MD	11	11
Algebra								
Expressions	14MD	14MD	38	38MD	29EM	29	20	20MD
Equations	17EM	17	41	41	32	32EM	23MD	23EM
Functions	20EM	20MD	44	44	35	35	26EM	26EM
Geometry								
Figures	23EM	23	14	14MD	38MD	38EM	29	29
Relations & Transformations	26	26	17MD	17	41EM	41MD	32	32MD
Coordinates	29	29EM	20EM	20EM	44	44	35EM	35MD
Measurement								
English & Metric Units	32	32EM	23	23	14MD	14EM	38EM	38
Length, Area & Volume	35	35MD	26EM	26EM	17EM	17	41	41MD
Angular Measure	38	38	29EM	29MD	20	20MD	44MD	44
Probability & Statistics								
Probability	41	41EM	32	32MD	23MD	23MD	14MD	14
Descriptive Statistics	44MD	44	35MD	35	26	26MD	17MD	17EM
Problem Solving								
Numbers								
Integers	3	3EM	3EM	3MD	3	3	3EM	3
Fractions	6MD	6MD	6	6	6	6	6MD	6EM
Percents	9EM	9EM	9	9	9MD	9MD	9	9EM
Decimals	12EM	12EM	12EM	12	12MD	12	12	12MD
Algebra								
Expressions	15	15	39	39EM	30	30EM	21MD	21EM
Equations	18	18MD	42MD	42	33EM	33MD	24	24
Functions	21EM	21MD	45	45MD	36	36MD	27	27
Geometry								
Figures	24	24	15EM	15MD	39EM	39	30EM	30MD
Relations & Transformations	27MD	27	18EM	18MD	42	42	33MD	33
Coordinates	30	30	21	21EM	45MD	45EM	36	36EM
Measurement								
English & Metric Units	33EM	33	24MD	24	15MD	15	39MD	39MD
Length, Area & Volume	36MD	36MD	27	27EM	18	18EM	42	42EM
Angular Measure	39	39	30MD	30	21EM	21	45EM	45MD
Probability & Statistics								
Probability	42EM	42EM	33	33	24	24MD	15EM	15
Descriptive Statistics	45MD	45	36MD	36EM	27EM	27EM	18	18

testing procedures. The administrators and teachers of most schools were highly cooperative and facilitated the process. One school, however, had to be dropped from the sample because efforts to implement participation were resisted. Two other schools were passed over because recent reorganizations at the district level brought their eighth-grade enrollments below 25 pupils, the minimum number required to obtain reasonably accurate estimates of school-level performance. In all cases, schools from the appropriate cells of the sample design were substituted.

On the basis of enrollment figures supplied by a school, a packet of testing materials was assembled for each class in the school. Each packet contained a sufficient number of pretests, answer sheets, and second-stage booklets for the pupils in the class. In each packet, the booklets from the eight second-stage forms were compiled in rotation within each level of difficulty. The packet also contained a copy of the test administration guide, the feedback form, the classroom transmittal form, and the transparent template.

NORC shipped the preassembled materials from Chicago to the participating schools in California. Unfortunately, the materials for one school failed to arrive on time. Because the principal of this school was reluctant to reschedule the testing sessions, we drew a replacement school from the appropriate cell of the sample design.

Once the materials arrived at a school, NORC field personnel met with the teachers at that school to distribute and explain the testing materials. They briefed the teachers on the purpose of the study and the procedures for administering the test and asked them to complete the feedback form after administering the test. NORC personnel returned to eight of the schools to observe the testing sessions.

4.4 Administration of the two-stage booklets in California

In October and November of 1987, the teachers administered the two-stage forms to eighth-grade students in their classrooms in split-session format, completing the pretest and second-stage test on different days in separate sessions. The teachers scored the pretest between the sessions and assigned each student to a second-stage booklet. At each level of difficulty, the assignments rotated within each classroom over the booklets from the eight forms. Approximately 700 pupils completed each of the eight forms, or about 230 per booklet.

The teachers reported very few difficulties with the testing procedure, but many felt that the main test was too long to be administered in one class period, especially at schools where the periods were less than 45 minutes long. Nonetheless, in the testing sessions monitored by NORC personnel, only a few students were unable to complete the test in the allotted time. In these classrooms the teachers quickly brought their classrooms to order and efficiently distributed the testing materials, leaving students with most of the period to work on the test.

In 28 of the 32 schools, the teachers scored the pretest with the transparent template and assigned each student a second-stage booklet. At two schools, teachers used a Scantron for scoring. At two other schools, the administrators asked NORC personnel to score the pretest and assign the booklets because they thought the activity would consume too much of their teachers' time. In this connection, six teachers, or about six percent of those who participated in the study, reported that scoring the pretest was too time-consuming. Nonetheless, nearly 70 percent of the teachers who completed the feedback form, reported that the pretest provided a helpful introduction to the main test.

4.5 Processing of the California data

After the testing sessions, the coordinators and teachers at each school prepared the testing materials for shipment to NORC, and NORC personnel arranged for the pickup of the materials at the schools and for their return to Chicago where the data-processing and analysis would be performed. The teachers and coordinators were instructed to bundle the materials by classroom and place the classroom transmittal form on top of each bundle (see Appendix G). This form identified the class and asked the teacher to identify special students (honors/advanced placement, special education, learning disabled, and ESL) who were members of their class. While most of the teachers carefully followed the procedure, some of the materials were returned without this identifying information. When this problem occurred, NORC personnel requested the school coordinators to supply class rosters identifying the special students.

Inspection of the testing materials also revealed that the booklet codes had not been entered on a small percentage of the answer sheets. In most cases, the problem was readily solved because the pupils and teachers had carefully followed directions and returned the materials with the answer sheets inside the corresponding booklets. In the relatively few cases where the booklets and answer sheets were separated, each answer sheet was assigned the booklet number that yielded the highest number-right score when keyed against all booklets. In this preliminary cleaning of the data, the clerical assistants also checked and corrected the school identification codes and inserted codes for special education, learning disabled, and ESL students on the answer sheets.

After this preliminary cleaning, the answer sheets were separated from the booklets, grouped by class, and collated with header sheets that uniquely identified the set of answer sheets from each classroom. The answer sheets were then machine-read and the student name-fields were cleaned. Of the 5,625 students who completed the second-stage forms, 514 were identified as special education, learning disabled, or ESL students. These students were eliminated from the item calibrations.

4.6 Scoring the Duplex instrument

As we mentioned in earlier chapters, the item-structure of the Duplex instrument is designed to provide two separate sets of scores at the pupil level, one set for the three mathematics proficiencies (Procedural Skills, Conceptual Understanding, and Problem Solving), the other for the five content areas (Number, Algebra, Geometry, Measurement, and Probability and Statistics). The process-proficiency scores aggregate over items in each column of the content-by-process design; the content scores aggregate over items in each row of the design. In these scores, the item responses of each pupil appear twice, once in the process scores, and once in the content scores.

Scores for the second-stage test were computed with the BILOG program of Mislevy and Bock (1983). For reporting purposes, the scores for each scale were rescaled to a mean of 250 and a standard deviation of 50 in the state. We discuss the results from these analyses in detail in Chapter 5.

The scales for the school-level scores were obtained through aggregation of items across booklets within each element or curricular objective of the content-by-process design. Each scale is thus represented by 24 items, one from each of the three booklets of the eight forms, including some common items that provide links among the booklets. The data for these calculations are simply the number of students who attempt each item within a school, and the number among these who respond correctly. We analyzed these data using a group-level model described by Mislevy (1983). As in the analyses of the pupil data, we assumed a three-parameter logistic model for the probability of a correct response as a function of the scale value for the school. The scores were then transformed to the metric of the individual-level scores on the basis of the results from an analyses of variance in which the background variables served to model between-school variation. We discuss the results from the group-level analysis in detail in Chapter 6.

4.7 Reports to the Schools

On the basis of Bock and Mislevy's (1988) evaluation of the potential uses and users of the information from the Duplex Design, we designed prototype reporting forms to address the special needs of pupils, teachers, principals, and superintendents. We describe these forms in Chapter 7. Copies of the forms were prepared with the TEX typesetting system and the results from the California study were overprinted with a laser printer. In March of 1988, NORC returned the student-level reports and classroom summaries to the California schools for distribution to the students and teachers. In April of 1988, it sent copies of the school-level report and the state-level distribution of pupil scores to the participating principals; copies of both of the state-level reports were returned to the district superintendents.

4.8 Summary

The field procedures of this study were carried out so as to resemble as closely as possible an operational assessment. They were developed first for the Illinois study, then perfected further for the California study. The main points of the protocol were:

1. The schools should be chosen in a stratified random sample based on five factors of the community background.

2. The request for cooperation of the school should come in a letter to the district superintendent from the chief state school officer, accompanied by a similar letter to the school principal from the director of the state education assessment program.

3. The Duplex instrument should be administrated as a two-stage test by classroom mathematics teachers, as instructed by NORC interviewers who visited the schools.

4. Materials for the testing should be delivered to and returned from the schools directly to NORC via United Parcel Service.

5. Computerized reports of the assessment results should be sent to the schools, teachers, and students who participated in the study, as well as to the respective district superintendents and to the state assessment program (see Chapter 7).

These procedures resulted in a high degree of cooperation from the schools and teachers involved, and they produced data of good quality and completeness for subsequent analysis. The Duplex instrument and its method of administration met quite adequately the conditions that might obtain in an operational state assessment program.

Chapter 5

Measurement at the student level

The distinctive feature of the Duplex Design is the provision for scoring the students on main content and proficiency dimensions of the subject matter. We have argued in Chapter 1 that scoring at this level is essential to insure a serious involvement of students in the outcome of the assessment, and to provide attainment data in a form that is most suitable for public discussion and secondary analysis. The more detailed school-level scoring, discussed in Chapter 6, is equally essential in a comprehensive assessment, but it is designed for measuring the highly specific objectives of learning that are the focus of curricular planning and management of instruction.

As we noted in Chapter 3, the main problem in obtaining student-level scores from assessment data is how to extract a sufficient amount of dependable information about the student from the relatively short forms that make up a matrix-sampled instrument. We proposed a solution to this problem based on two technical innovations—the conjoint scoring of content and proficiencies, in which each item response does "double duty" by contributing to two scales—the use of two-stage testing to obtain more information per item response from students at different levels of proficiency. Neither of these procedures is a necessary part of the Duplex Design—student-level scores could be computed independently for content or process dimensions, and the students' responses could be obtained with a conventional one-stage test. But their successful incorporation into the testing and scoring procedures will appreciably improve the cost-benefit of the Duplex methodology.

The Illinois field trial enabled us to work out efficient implementation of these procedures, and the California trial, which we report here, provided an evaluation of them in revised and improved form. The object of the latter trial was to answer the following questions:

1. Is the two-stage testing procedure practical for an assessment program that depends on local school personnel to administer the instrument?

2. Are the student-level content and proficiency scores of the Duplex Design sufficiently reliable for use in guidance and certification?

3. Do the student-level scores have good technical properties for use in secondary research?

4. Does the information gain from two-stage testing justify the additional complexity of test administration and analysis?

We attempt to answer these questions in the remaining sections of this chapter.

5.1 Feasibility of two-stage testing

As mentioned earlier, two-stage testing has seen little practical use, probably because it is considered too cumbersome to administer. This has not been our experience in either the Illinois or the California field trial: in our implementation of the procedure, neither the students who took the test nor the teachers who administered it reported any special difficulties due to the method of test administration. To obtain information on the teachers' reactions we included with the school materials a questionnaire about the test administration; we also had NORC field personnel observe the testing in 8 of the 32 schools. In addition, we judged the extent to which teachers followed instructions by examining their scoring of the pretest and their assignment of students to the second-stage forms.

The two-stage testing implementation described in Chapter 4 was used in the California field trial. The classroom teachers administered the 15-item pretest on a day preceding the second-stage test.

The instrument made use of separate answer-sheets on which students recorded their responses, first to the pretest, and later to the second-stage test items. Between the two stages, the teacher or an assistant scored the pretest using the supplied transparent template and, following the guide on the template, assigned each student to a Difficult, Medium, or Easy second-stage booklet. In preparation for returning the booklets to the students for the second testing, the teacher marked the assignment on the student's answer sheet and inserted the sheet beneath the cover of the second-stage booklet, which was trimmed so that the student's name would show.

5.1.1 Student motivation

As in all types of cognitive testing, the two-stage procedure is most effective when the students are motivated to perform to the best of their ability. Each student is then accurately classified by the first-stage test and receives a second-stage booklet matched to his or her level of proficiency. There is a gain in reliability because at the second stage the examinee is presented with items that are informative at his or her level of attainment.

This gain may be compromised if for any reason the first-stage score does not reflect the student's general ability to answer the second-stage items. The loss is tolerable if the misassignment of the student is limited to one difficulty step (*e.g.*, assignment to the Medium test booklet instead of the Easy booklet, or Difficult instead of Medium). The effect on the second-stage score is reduced by the considerable overlap in item difficulty between adjacent booklets (due in part to the linking items).

But if by mistake a student is assigned an Easy booklet instead of a Difficult booklet, or vice-versa, some error would result if the student answered all of the items correctly or incorrectly in the inappropriate second-stage booklet. When the scoring procedure combines the information from the first-stage classification and the responses to items in the second stage, it would assign a lower score than if the student had been mistakenly assigned to the Medium booklet and answered all the items correctly. A converse error would result if a student who belonged in the Easy group were assigned to the Difficult group and answered all items at that level.

The effect would not be much different, however, from that of a one-stage test to which the examinee responded to a subset of items in a manner inconsistent with his or her ability. There is no reason to think that a two-stage procedure employing a small number of overlapping second-stage forms would be less robust than a comparable one-stage test; in either case the validity of the test scores will be influenced by the quality of the student's test-taking efforts. The accuracy of both types of testing will benefit from the student's personal stake in the outcome; but two-stage testing will also save testing time.

5.1.2 Administration: cooperation of the teachers

The implementation of the two-stage testing procedure in the field trials relies on the cooperation of teachers. They are expected to score the pretests and assign the second-stage booklets correctly. By and large, the Illinois and California teachers responded well to these tasks and, with a few exceptions, assigned the second-stage test booklets accurately.

To check the teachers' assignments, we rescored the pretest for each pupil from the optically scanned answer sheets and compared it with the group assigned by the teacher.[1] Three schools were omitted from this verification. In two of these schools, NORC personnel scored the pretests and assigned the second-stage booklets; at the other, the scoring templates were misplaced and teachers assigned the booklets on the basis of the student's previous classroom performance. The results for the 29 remaining schools are presented in Table 5.1.

The diagonal elements of Table 5.1 show the number and percent of teacher assignments that agree with our computer-calculated scores. They reveal that the teachers assigned the booklets with about equal accuracy at the three levels of difficulty. They also show that more than 95 percent, or 3,907, of the 4,108 pupils received test booklets that were in accord with their pretest performance.

The off-diagonal elements of the assignment table show the number and percent of teacher errors. Most of the errors, 79 percent, are

[1] Students with 0–6 items correct on the pretest were assigned to the Easy booklet, 7–10 to the Medium booklet, and 11–15 to the Difficult booklet.

TABLE 5.1
Number and percent of second-stage booklets correctly and incorrectly assigned

	Correct Booklet			
Booklet Assigned	1	2	3	Total
1	1552	35	11	1604
	94.5%	2.4%	1.7%	
2	66	2367	14	1447
	4.0%	94.5%	1.4%	
3	25	44	988	1057
	1.5%	3.1%	96.9%	
Total	1643	1446	1019	4108

at adjacent levels of difficulty; *e.g.*, a pupil who should have received an Easy booklet, was assigned a Medium booklet instead. The fact that the teachers were told to substitute booklets from adjacent levels of difficulty if they ran out of booklets at the appropriate level could account for most of these errors. The misassignments that are off by two levels of difficulty are harder to explain. Three of the 107 teachers in the California study are responsible for all but two of these errors, displayed in the upper right- and lower left-hand corners of the assignment table. These same three teachers are also responsible for more than 46 percent of the total errors made by the teachers. Whether these teachers misplaced or misaligned the templates, misunderstood the directions, or simply refused to follow the procedure is not entirely clear. Their errors are about equally divided among the off-diagonal elements of the assignment table.

In addition to the correct assignment of the second-stage forms, the teachers were also expected to use the forms in the order they had been packaged by NORC. An assumption of matrix-sampling methodology is that any student has equal probability of receiving any one of the assessment forms. It is standard procedure in assessment to pack the forms in rotation *beginning successive packets at a different point in the series*, and to require the teachers to distribute them in rotation to students in the classroom. This insures that the unused portions of the packets within a school are not biased toward the later forms in the series and that all forms will be used about equally often in the total sample. To check on this condition, we looked at the distribution of booklet assignments across the eight forms of the second stage. We found all of the distributions to be consistent with the conclusion that the teachers assigned the booklets in rotation, except for small differences that can be attributed to random variation in the procedure used to package and assign the second-stage booklets.

The remaining source of information about the teachers' role in the test administration was their responses to the brief questionnaire shown in Appendix F of this report. On the whole, the teachers responded favorably concerning the testing procedure. They reported that the written instructions for administering the test were easy to follow and the transparent scoring templates, easy to use. Although a small percentage (6%) complained that the procedure was too time-consuming, more than 69 percent of the 72 teachers who completed the feedback form (71% of the teachers in the study) reported that the pretest provided a helpful introduction to the main test. It familiarized their pupils with the testing procedure and expedited the administration of the second-stage test. Others felt that the procedure reduced the frustration level of low-ability pupils who ordinarily perform poorly on standardized tests. More than 65 percent of the teachers who completed the form believed that the pretest properly identified the ability level of all or most of their students; about 21 percent reported that it failed to do so, but our analysis of the pretest scores in Section 5.2.2 shows this is not the case.

Most of the latter teachers complained that the pretest, as well as the main test, was too "wordy" for their limited-English-speaking students—that it measured these students' ability to understand English rather than mathematics. Apparently these teachers did not know that the Duplex instrument was not intended to assess the mathematics ability of limited-English-speaking students. To avoid singling them out as special students, ESL students were given the opportunity to participate in the study if they were enrolled in regular mathematics classes. They were identified from class rosters, however, and their responses were not used in the item calibrations, nor were their scores used in the computation of class, school, and state statistics.

5.2 Item analysis in the California data

We now turn to the item analysis required in scoring California test results at the student level. After describing the item response model and the method of calibrating the items of the instrument, we examine the item-parameter estimates from the scaling of the content and process categories. We then apply the item-parameter estimates in computing the scale scores that are used to report the mathematics attainments of the individual students. Finally, we carry out an information analysis of these scales, including estimation of their reliabilities in the California 8th-grade population, and evaluate the information gains due to the two-stage testing procedure.

5.2.1 The IRT model

A fundamental premise of item response theory is that an individual's cognitive proficiency is not directly observable in test performance, but must be inferred from his or her pattern of responses to the test items. The item responses, in turn, are assumed to be stochastically related to the proficiency variable; that is, they contain unrelated random variation, as well as variation determined by the underlying variable. To infer the proficiency level of the individual from the observed response pattern in the presence of this random error requires a correct representation of the probability of the pattern as a function of the underlying variable. Typically, this representation is expressed mathematically in terms of an *item response function*. These functions are statistical descriptions of what happens when an examinee encounters given items on the test. They express the probability of the response to the item as a function of the examinee's *position* on the proficiency continuum and certain *parameters* that describe the operating characteristics of the item in the test context.

There are now a number of response functions available for use with binary (right-wrong) scored items (see Lord, 1980). The most widely used are the so-called logistic models, and the most general of these is the three-parameter logistic, or 3PL, model. The 3PL model defines the probability of a correct response, $P_j(\theta)$, to item j as a function of three item parameters, a_j, b_j and c_j, and the proficiency level, θ, as follows:[2]

$$P_j(\theta) = P(x_j = 1|\theta) = c_j + \frac{1 - c_j}{1 + e^{-1.7a_j(\theta - b_j)}}, \qquad (5.1)$$

where

$$x_j = \begin{cases} 1 & \text{if the response to item } j \text{ is correct.} \\ 0 & \text{otherwise.} \end{cases}$$

The probability of an incorrect response is the complement, $P(x_j = 0|\theta) = 1 - P_j(\theta)$.

When (5.1) is plotted as a function of θ, the result is a curve such as one of those shown in Figure 3.2 or Figure 5.3. These plots are called *item trace lines*, or *item characteristic curves*. At any given value of theta, the probability of a correct response is indicated by the height of the item characteristic curve at that value. The dotted line in Figure 3.2 illustrates this relationship. They show that a person at scale point −0.6 has probabilities 0.99, 0.81, 0.42, 0.20, 0.08, and 0.01 of answering the successive items correctly. The shapes of the curves, and thus the response probabilities, reflect the values of their item parameters.

The c_j parameter in (5.1) corresponds to the lower asymptote of the curve. It indicates the probability of chance success on an item when an individual is totally lacking in ability and resorts to random guessing. For multiple-choice tests, it is common practice to estimate an unique c_j for each item in the instrument because its value is influenced by the number and content of the response alternatives, both of which may vary from one item to another. The tendency of examinees to make "blind" and "educated" guesses also affects the estimate of c_j. A value of c_j equal to the reciprocal of the number of response alternatives will result if an examinee who does not know

[2] The constant e in this expression is the base of natural logarithms; $e = 2.718...$

the answer elects to mark at random rather than leave the answer sheet blank. Educated guesses in which the examinee narrows down the number of plausible alternatives contribute to higher values of c_j. For example, the item shown in (a) of Figure 5.3 has a value of c_j equal to 0.21. For the curve shown in (b), c_j equals 0.106.

The b_j parameter of (5.1) indicates the location of the response function along the proficiency continuum. It reflects the difficulty of the item and is often referred to as the *threshold* parameter, although the term *location* parameter is more apposite and will be used here. The curve for Item 2 shows that when $c_j = 0$ the inflection point will be at the value of θ where the probability of a correct response to the item is 0.5. Item 1 shows that when $c_j \neq 0$, b_j will be at the point along the proficiency continuum where the probability of a correct response is equal to $(1 + c_j)/2$. The dotted lines in the figure illustrate these relationships. They also show that Item 2 is generally more difficult than Item 1; *i.e.*, the value of its b_j parameter is located further to the right on the proficiency continuum.

The third item parameter, a_j, is called the item *slope*. It is proportional to the rate of change of the item response function at the inflection point and is a measure of the discriminating power of the item. Its value is related to the correlation between the responses to the item and the underlying trait. Items with steeper slopes have higher item-trait correlations and are more discriminating than items with shallower slopes. When the scale scores of the examinees are computed, responses to items with higher discriminating power have more weight in determining the score than those with lower discriminating power.

In applications where guessing is discouraged or the test is composed of free-response items, the IRT model can sometimes be simplified by setting the c_j parameter to zero. In the logistic case, the resulting function is referred to as the "two-parameter" logistic, or 2PL, model. If there are no guessing effects and the items have been selected to have uniform discriminating power, the model can be simplified still further by assuming the a_j parameter of all the items in each subtest of the instrument to be equal. This most specialized of the logistic IRT models is referred to as the "Rasch", "one-parameter", or 1PL model.

The question of which model is appropriate in a particular application can be answered empirically by examining, in a large sample of data, the "goodness-of-fit" of the corresponding response functions. But in matrix-sampling designs adequate fit of the 1PL model is unlikely because there are too many items to permit the imposition of uniform discriminating power in all forms and scales. The 2PL model is a more credible possibility, for although the items are in the multiple-choice format and subject to guessing effects, the instructions to the students (see Chapter 4) clearly discouraged guessing. Nevertheless, rather than prejudge the matter, we will examine the fit of all three of the binary logistic models.

Given the fitted response functions for each item of a particular scale, the probability of the pattern of correct and incorrect responses can be computed on the assumption of *conditional* independence—namely, that the item responses are stochastically independent conditional on θ. Since the joint probability of independent events is the product of their separate probabilities, the probability of an answer pattern, $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, of n items may be expressed as

$$P(\boldsymbol{x}|\theta) = \prod_{j=1}^{n} [P_j(\theta)]^{x_j} [1 - P_j(\theta)]^{1-x_j} \qquad (5.2)$$

where $\prod$ represents the continued product of these terms. This formula, which may rightly be called a fundamental law of item response theory, was introduced by Lawley in 1943.

5.2.2 Choosing the item response model: analysis of the first-stage test (pretest)

Because responses to the separate second-stage test booklets do not represent the full range of variation in the population, they are not very useful in evaluating the fit of the assumed item response model. In the present study, these data are even less suitable for this purpose because the number of students who have responded to the items of each of the twenty-four test booklets is relatively small (in the neighborhood of 200, as opposed to 500 or more desirable for an adequate evaluation of fit). We can, however, analyze in detail the pretest items, where the number of cases is large ($N = 5{,}023$) and the responses represent the complete range of 8th-grade students.

To estimate parameters of the models, we used the marginal maximum likelihood (MML) method of Bock & Aitkin (1981), as implemented in the BILOG program of Mislevy & Bock (1983). It is the only statistically rigorous method available for fitting all three of the binary logistic models—indeed, it is valid for all functional item response models, whether the response is scored dichotomously or in multiple categories. A fundamental assumption of the MML method is that the respondents are drawn from some population, unequivocally the case in large-scale assessment applications. On this assumption, the MML method integrates the right member of (5.2) over the population distribution of θ to obtain the unconditional, or marginal, probability of each of the patterns of correct and incorrect responses that occur in the sample. It then invokes the maximum likelihood principle and assigns values to the item parameters that maximize the product of these sample probabilities.

The form of the population distribution can be either assumed or, if the sample size is large enough, estimated along with the item parameters (see Mislevy, 1984). When applied to assessment data for public schools where the children are not specially selected, it is reasonable to assume a normal distribution for the measure in question. Small discrepancies between the actual and assumed distribution have little effect on the estimation of the item parameters and can be safely ignored. For the probability sample of the present study, for example, it is quite satisfactory to assume a normal population distribution of overall mathematics ability as measured by the pretest, which was administered to all the testable students in the participating schools. Because the sample size for the pretest is in excess of 5,000 students, we were able to check the assumption of normality by estimating the latent distribution. We see in Figure 5.1 that the empirical distribution conforms closely enough to the normal to justify this assumption of MML estimation of the pretest item parameters.

Given a specified population distribution, the MML method provides statistical tests of the fit of 1PL model relative to that of the 2PL model, and of the 2PL model relative to the 3PL model. These tests are based on the difference in the logarithms of the maximum

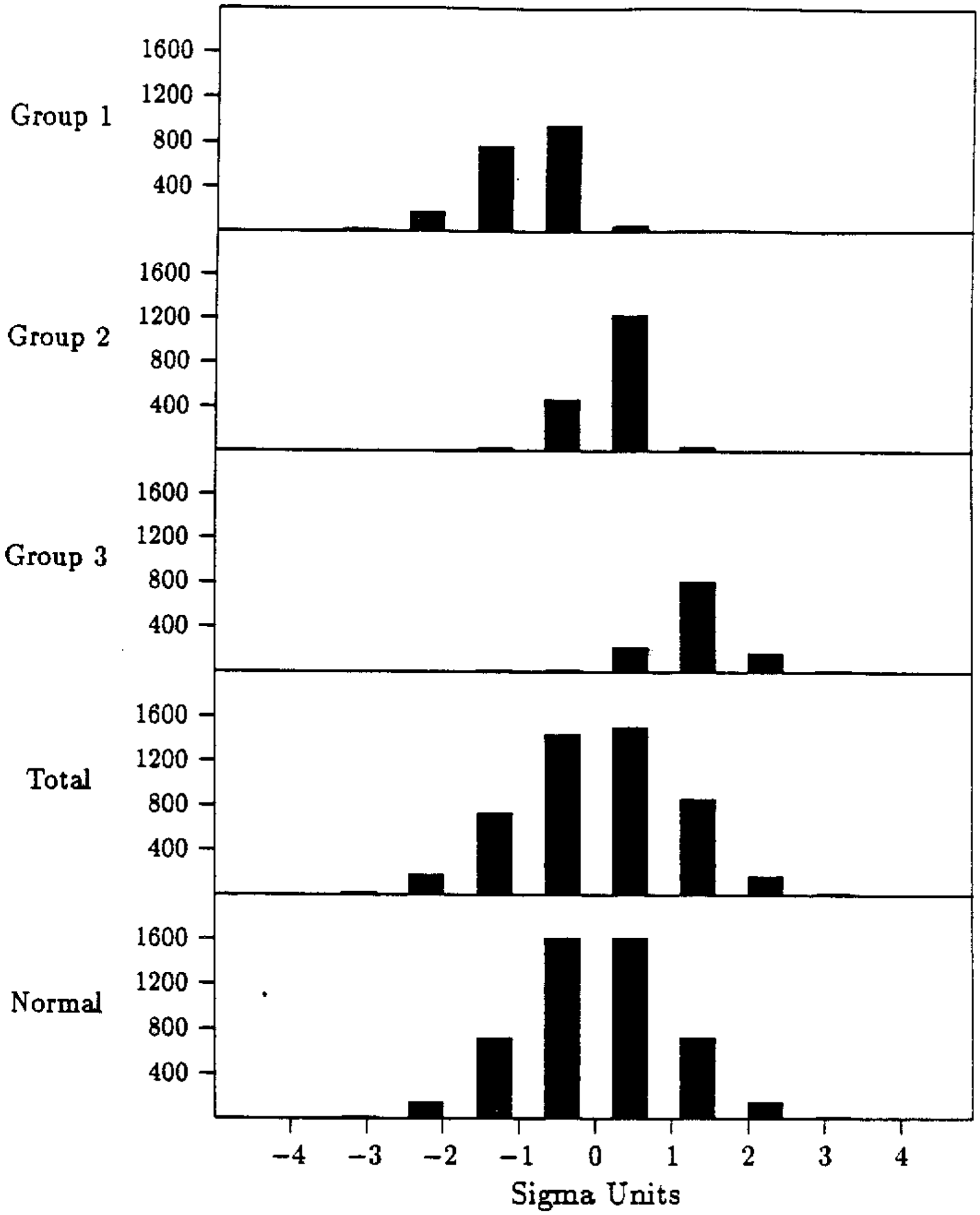


Figure 5.1. Estimated latent distributions based on the pretest data. The total for groups 1, 2, and 3 conforms well to the normal distribution (bottom).

TABLE 5.2
Likelihood-ratio tests of the fit
of the 2PL model relative to the 3PL,
and of the 1PL model relative to the 2PL

Model	−2 log likelihood	Difference	d.f.	Probability
1PL	15589.45			
		314.39	15	<.00001
2PL	15275.06			
		35.15	35	<.005
3PL	15239.91			

marginal likelihoods obtained in fitting the respective model. In large samples, twice the positive differences are distributed as chi-square variables on degrees of freedom equal to the number of items. For the pretest, calculated from a random 1,000 cases from the full sample, the maxima of the log likelihoods and the corresponding differences for the models are shown in Table 5.2. As expected, the 1PL model is grossly inappropriate for these data. The 2PL model does much better, but technically is a significantly poorer fit than the 3PL model.

The difference chi-squares in Table 5.2 bear only on the *relative* fit of the 2PL model versus the 1PL, and the 3PL versus the 2PL. They do not speak to the question of the *absolute* fit of the 3PL model. Indeed, for tests with relatively small numbers of items, such as the 15-item pretest, no entirely satisfactory method of evaluating absolute fit is available. But a graphic impression of the fit is conveyed by a plot of the posterior probability of a correct response, based on the data, compared to the probability predicted by the fitted model at selected points on the latent continuum. Plots of this type are provided by the BILOG 3.0 program. For the 1,000 case sample, plots showing an average-fitting item and the worst-fitting item are presented in Figure 5.2. Each plot includes plus-or-minus two standard deviation credible intervals for the probability at selected points. With each plot is given the root-mean-square sum of the standardized residual weighted by the expected number of respondents at each point (points with expected numbers of respondents less than five are not included.) In tests with indefinitely large numbers of items, this quantity becomes equal to a chi-square statistic divided by its degrees of freedom. As such it might be considered to raise a question about the fit when it becomes greater than 2.0. That only one item [shown in Figure 5.2(b)] among the 15 exceeded this value reinforces the impression from the plots that the absolute fit of the pretest items to the 3PL model is reasonably good.

From the plot of the poor-fitting Item No. 8 in Figure 5.2, it is apparent that the students in the -0.30 to -0.90 range are performing better than expected relative to those in the 0.30 to 0.90 range. To fit these data well, the response function would need two inflection points—a shape that the 3PL model cannot accommodate. Inspection of the item content reveals the likely source of this effect:

> Mary has three feet of candy cane that she wants to divide equally among six friends. How long will each piece be?
>
> A. $\frac{1}{2}$ inch
> B. 6 inches
> C. 1 foot
> D. 18 inches
> E. 12 feet

Undoubtedly, the students in the medium-low range are picking up on the "six" in the item stem and choosing alternative B without attempting to understand the problem. As a result, the item responses are a mixture of two strategies, one of which is wrong but fortuitously gives the correct answer, while the other is fully correct. This is a common fault in item writing; it leads to items such as No. 8, with lower discriminating power (slope = 0.790) and poor fit.

The practical effect of the marginally poor fit of Item 8 on the scoring of the pretest is small, however. Considering the excellent fit of the remaining items, we have ample justification for choosing the 3PL model for our analysis of the California data.

Having verified the form of the latent distribution and chosen the IRT model, we are now in a position to estimate the parameters of the pretest items by the general MML method. The resulting estimates as obtained by the BILOG program from the sample of 1,000 cases

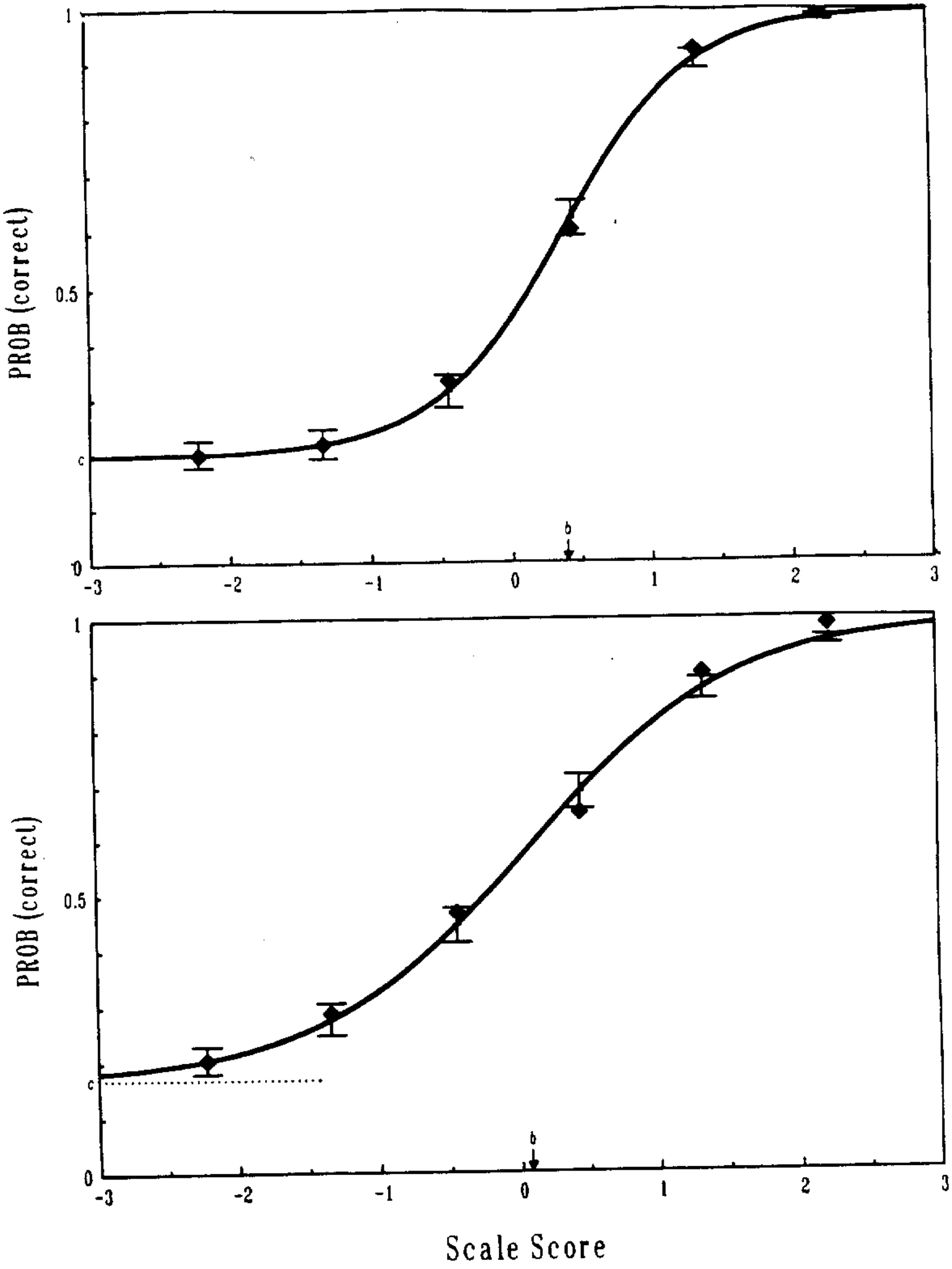


Figure 5.2. Fitted 3PL response functions for pretest item:
No. 4 ($a = 1.267$, $b = 0.391$, $c = 0.197$),
No. 8 ($a = 0.790$, $b = -0.070$, $c = 0.169$).

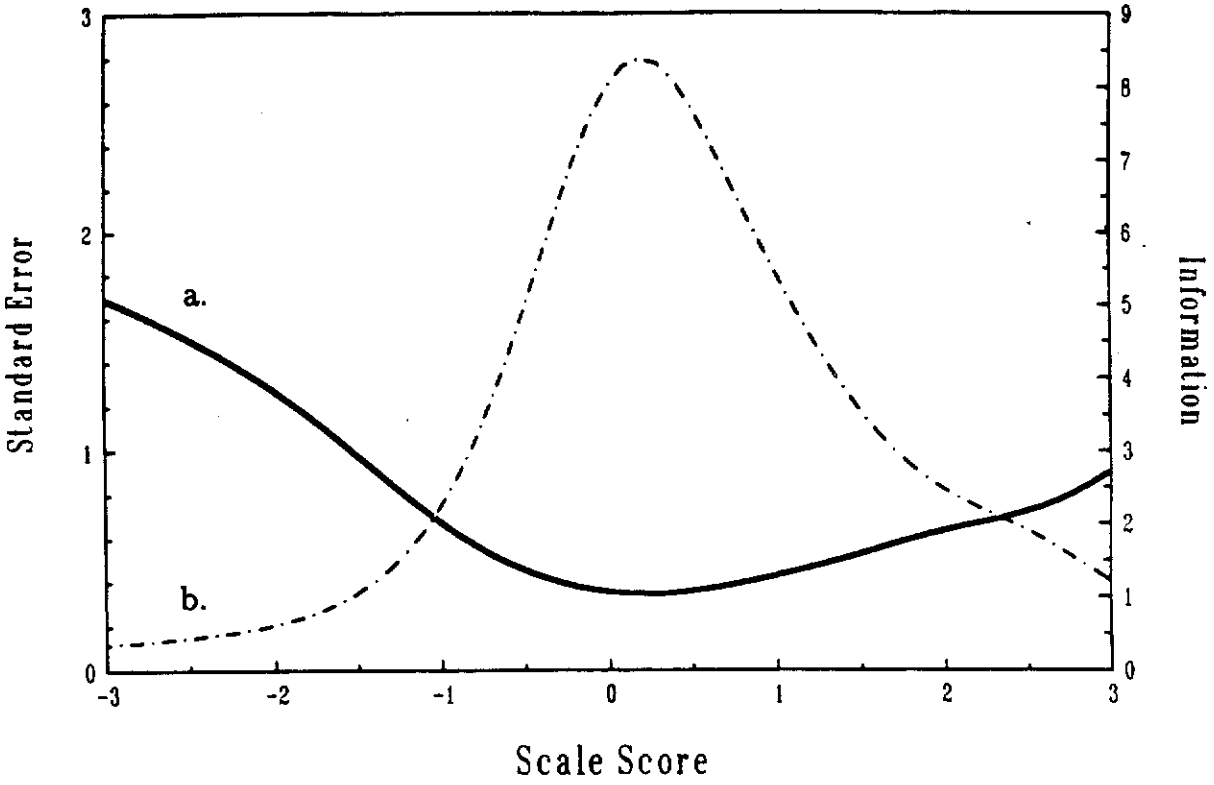


Figure 5.3: The standard error (a) and information (b) curves for the 15-item pretest.

sample of *pretest* data. We justify this assumption on the grounds that all of the scales are rather strongly correlated with the overall mathematics proficiency as estimated by the pretest. Moreover, we know from experience with MML estimation of item parameters that it is only the provisional location of the prior distributions of the three groups that is important, and the pretest data locates them with sufficient accuracy.[3] The link items then provide the adjustments to the provisional locators as required to position correctly the latent distribution for each scale of the second-stage test.

To estimate the latent distributions for the three pretest groups, we make use of the provision in the BILOG program for obtaining discrete point distributions along with the calculation of expected *a posteriori* (EAP) scale scores for the students in the respective groups. These are the estimated distributions for the second-stage groups shown in Figure 5.4. Using this partition of the latent distribution, we fitted the 3PL models for the content and process scales of each form and booklet of the Duplex instrument.

5.2.4 Linking the scales

In fitting the link items common to the Easy and Medium, and to Medium and Difficult test booklets, we make additional special provisions so that the curve for each item will have the same shape in each group. In so doing, we bring the parameter estimates from the three test booklets on to a common scale. For this purpose, we first assume that only the *location* of the corresponding response functions will vary between the two booklets in which each link item appears. Both the discriminating power and the lower asymptote parameter are constrained to have the same value in the two analyses containing the same link items. This procedure ıs justified on the grounds that the selection of the students by level of achievement on the pretest should affect only the absolute positions of the items on the latent con-

[3] Alternatively, we would have enough data in the sample to estimate a *common* latent distribution for all eight of the second-stage forms simultaneously. Then we would not have to use the pretest information for this purpose. Unfortunately, at the time of this writing we did not have a computer program suitable for this type of multiple-group estimation.

tinuum; it should not affect their discriminating powers or the lower asymptote. For convenience, and also because the slopes are better estimated toward the center of the distribution, we have imposed these constraints by setting the slope and asymptote values for the link items of the Easy and Difficult booklets equal to the corresponding unconstrained estimates obtained from the Medium booklet. The BILOG program has provisions for setting item parameters to specified values by manipulating the item prior distributions. The non-link items are, of course, estimated separately, without constraints, in the data for the respective booklets and forms.

We illustrate the results of these analyses in the California data for one of the process categories, Procedural Skills, and one of the content categories, Algebra. Under the assumed restrictions on the slope and asymptote parameters, only the location parameters for items in the Easy and Difficult booklets need to be adjusted to bring all the estimates to the same scale. The required adjustments are the arithmetic means of the differences between the location estimates of the link items for the Easy and Medium booklets, and for the Medium and Difficult booklets, as shown in Table 5.4. The location of the latent distribution for each of the three groups, as determined from that of the pretest, is sufficiently accurate for the Procedural Skills and Algebra scales that these differences are in the neighborhood of zero; some are positive and some, negative. That they are somewhat variable is the fault of the rather small size of the second-stage form and booklet samples. In full state-wide implementation of a Duplex Design, these sample sizes would be much larger and the difference for the individual link items would be more consistent.

Effects of sampling variation notwithstanding, the link estimates for Algebra Item 21 in Form 1 and Procedural Skills Item 31 in Form 3 seem out of line. The latter item is excessively difficult and should not have been used for linking. The former has two possible solution strategies, one of which may have facilitated the Group 1 students. If similar anomalies occur among link items generally, a robust estimate of the mean value for the adjustment should be employed. Even in these data, however, their values are about the same for the two forms, as they should be for randomly parallel forms. We have therefore averaged the differences from both Form 1 and Form 3 to obtain the

TABLE 5.4
Location adjustments between test booklets,
based on separately estimated link-item location

	Booklets					
	Easy and Medium (EM)			Difficult and Medium (DM)		
	Form	Item	Difference (E−M)	Form	Item	Difference (D−M)
Procedural Skills	1	7	−0.032	1	1	0.318
	1	10	0.054	1	4	0.132
	1	31	−0.145	1	16	0.122
	1	34	−0.168	1	37	0.433
	3	13	−0.147	3	41	0.292
	3	25	−0.106	3	10	−0.207
	3	31	0.903	3	28	0.280
	3	37	0.003	3	34	0.603
	3	43	−0.385			
Adjustment (mean)			−0.002			0.247
Algebra	1	17	−0.466	1	14	0.416
	1	20	−0.449	1	16	0.139
	1	21	−1.256	3	42	0.333
	3	37	−0.342			
	3	43	−0.411			
Adjustment (mean)			−0.585			0.296

linking adjustments for the two forms as shown in Table 5.4.

At this stage in the calculations, the values for the Medium booklet define the origin and unit of measurement. (Later the origin and unit for each scale will be adjusted so that the student-level scores have mean 0.0 and standard deviation 1.0 in the California sample; still later, for reporting purposes, the mean will be moved to 250 and the standard deviation to 50.) Finally, we apply these adjustments to the provisionally estimated location parameters and, rescaling to mean 0 and standard deviation 1 in the latent distributions, obtain the item-parameter estimates shown in Table 5.5. When scaled in this way, we say that the item-parameters estimates are in the "0,1" metric. It is a convenient choice of scale for purposes of comparing item parameters estimated for different populations.

The estimated item parameters shown in Tables 5.5(a) and 5.5(b) are in the order in which the items appear in each booklet with link items indicated by ME or MD, as the case may be. The values in this table are in the form required for computing student-level scores on a common scale, regardless which second-stage booklet any particular student was assigned. That is, the parameters correspond to the items as they appear in the order of the Procedural Skills scale and the Algebra scale of the three test booklets in Forms 1 and 3. The values for each link item are represented twice, once for each booklet in which they appear; the values of their locations are the averages of the estimates from the two booklets. Similar tables apply to the other content and process scales and the other six second-stage forms.

We see from the estimated locations of the items that the difficulty levels of the three tests booklets are positioned more or less as intended, except that the items as a whole are more difficult than they should be for this population. This is evident in Tables 5.5 (a) and (b) from the relatively few items with locations below zero. Since the mean score for the population has been set at zero, the test would be more informative if the distribution of item locations were more centered on zero. But the lower than expected proficiency levels in the California students relative to the Illinois students, which we mentioned above, misled us in our choice of items for the revised instrument. Even with well-positioned items, the appearance of the link items in two booklets results in the overlap of item difficulties

TABLE 5.5
Item-parameter estimates for two scales
of the second-stage test (a) Form 1

Procedural Skills			
Item	Slope	Location	Asymptote
1E	0.685	−1.420	0.183
4E	1.056	0.366	0.112
7EM	1.113	0.542	0.195
10EM	1.671	0.094	0.141
13E	1.000	0.630	0.176
16E	0.933	−1.626	0.166
19E	0.750	0.080	0.201
22E	0.811	−0.739	0.202
25E	1.007	−0.669	0.188
28E	0.748	−0.437	0.151
31EM	1.219	0.465	0.142
34EM	1.066	1.027	0.229
37E	0.933	−0.113	0.159
40E	0.897	2.492	0.133
43E	1.602	−0.358	0.138
1MD	1.021	0.857	0.169
4MD	1.600	0.605	0.132
7ME	1.113	0.542	0.195
10ME	1.671	0.094	0.141
13M	1.277	0.231	0.155
16MD	1.667	0.600	0.112
19M	0.843	0.709	0.182
22M	0.849	0.238	0.169
25M	0.671	0.660	0.180
28M	0.793	0.814	0.192
31ME	1.219	0.465	0.142
34ME	1.066	1.027	0.229
37MD	0.937	0.725	0.167
40M	0.915	1.812	0.144
43M	0.850	0.198	0.180
1DM	1.021	0.857	0.169
4DM	1.600	0.605	0.132
7D	1.376	1.175	0.115
10D	0.693	2.607	0.176
13D	1.492	1.017	0.150
16DM	1.667	0.600	0.112
19D	1.717	1.976	0.116
22D	1.131	1.094	0.140
25D	1.179	2.140	0.076
28D	0.972	1.103	0.134
31D	1.213	1.310	0.190
34D	0.963	2.195	0.139
37DM	0.937	0.725	0.167
40D	1.278	1.499	0.147
43D	0.949	2.263	0.102

Algebra			
Item	Slope	Location	Asymptote
13E	1.259	0.560	0.130
14E	0.707	0.128	0.189
15E	1.106	0.656	0.236
16E	1.063	−1.213	0.160
17EM	0.972	−0.433	0.164
18E	0.771	0.650	0.216
19E	0.819	0.310	0.176
20EM	0.964	1.626	0.095
21EM	1.077	0.011	0.145
13M	1.608	0.069	0.136
14MD	1.739	0.474	0.098
15M	0.958	2.154	0.146
16MD	1.417	0.478	0.115
17ME	0.972	−0.433	0.164
18M	1.249	0.753	0.123
19M	1.162	0.418	0.159
20ME	0.964	1.626	0.095
21ME	1.077	0.011	0.145
13D	1.756	0.902	0.128
14DM	1.739	0.474	0.098
15D	1.677	1.396	0.189
16DM	1.417	0.478	0.115
17D	1.592	2.101	0.086
18D	0.930	1.186	0.144
19D	1.758	1.848	0.106
20D	0.523	3.025	0.221
21D	1.796	1.032	0.119

TABLE 5.5 (continued)
Item-parameter estimates for two scales
of the second-stage test (b) Form 3

Procedural Skills			
Item	Slope	Location	Asymptote
1E	0.482	−0.826	0.214
4E	1.554	−0.398	0.113
7E	0.702	−0.472	0.216
10E	0.785	−1.133	0.180
13EM	0.938	0.120	0.152
16E	1.093	−0.505	0.181
19E	0.927	0.721	0.181
22E	0.762	0.048	0.183
25EM	0.802	−0.217	0.161
28E	1.233	−0.437	0.167
31EM	0.780	1.749	0.226
34E	0.994	0.185	0.177
37EM	0.699	0.044	0.173
40E	0.766	−1.097	0.173
43EM	1.334	0.063	0.157
1M	0.574	−0.198	0.172
4MD	1.868	0.559	0.106
7M	1.584	0.557	0.107
10MD	0.640	0.331	0.174
13ME	0.938	0.120	0.152
16M	1.490	0.673	0.152
19M	1.337	0.736	0.184
22M	0.986	1.681	0.144
25ME	0.802	−0.217	0.161
28MD	1.259	1.446	0.137
31ME	0.780	1.749	0.226
34MD	1.054	0.916	0.166
37ME	0.699	0.044	0.173
40M	1.138	0.483	0.151
43ME	1.334	0.063	0.157
1D	1.505	1.451	0.214
4DM	1.868	0.559	0.106
7D	1.013	0.896	0.119
10DM	0.640	0.331	0.174
13D	0.413	3.496	0.182
16D	0.626	2.010	0.147
19D	1.205	2.125	0.106
22D	1.526	1.125	0.190
25D	1.340	1.567	0.108
28DM	1.259	1.446	1.137
31D	1.428	1.038	0.142
34DM	1.054	0.916	0.166
37D	0.515	4.544	0.113
40D	1.286	1.970	0.111
43D	0.716	3.437	0.120

Algebra			
Item	Slope	Location	Asymptote
37EM	1.047	0.115	0.157
38E	1.374	0.080	0.128
39E	0.684	−0.564	0.181
40E	0.697	−0.737	0.180
41E	1.158	−0.698	0.151
42E	0.826	0.019	0.204
43EM	1.190	0.233	0.149
44E	0.842	2.841	0.210
45E	0.970	−0.308	0.157
37EM	1.047	0.115	0.157
38M	0.609	0.464	0.208
39M	0.962	0.654	0.170
40M	1.507	0.317	0.124
41M	0.987	0.517	0.151
42MD	0.749	1.070	0.198
43EM	1.190	0.233	0.149
44M	0.854	1.542	0.148
45M	1.475	0.346	0.122
37D	0.515	4.331	0.104
38D	1.046	0.185	0.156
39D	1.320	0.441	0.145
40D	1.040	1.981	0.111
41D	1.142	2.126	0.125
42DM	0.749	1.070	0.198
43D	0.559	3.880	0.122
44D	1.070	2.030	0.137
45D	0.841	1.308	0.129

that is apparent in the estimated location parameters for the three booklets.

As indicated schematically in Figure 2.1, this overlap in the item difficulties is necessary in the scoring of second-stage tests. But we now realize that we were too conservative in requiring *four* link items per scale, per pair of booklets, per form. Because the forms are randomly parallel, the same adjustment constants apply in all forms and need to be estimated only once. Thus, one link item per scale per booklet per form would have supplied eight items on which to base the common linking adjustment, more than enough for an accurate linking. With only a one item overlap, the item difficulties in the second-stage booklets could have been more accurately placed and the student-level scores made more reliable. The effects of the assignment of items to second-stage booklets are discussed in Section 5.4.1 on test information.

Despite these problems with the revised form of the instrument, the relatively high discriminating ability of the items, which is apparent in the tables from the large proportion of items with slopes greater than 1.0, gives every prospect that the scores for the student-level content and process profiles, as well as the overall mathematics attainment score, will have good reliability for the California population. We present estimates of these reliabilities when we examine the information properties of the scales in Section 5.5.

5.3 Scoring the Duplex instrument

Once the estimates for the item parameters are known, the computation of the student-level scale scores is straightforward. Of the three methods of scoring that the BILOG program offers—Maximum likelihood, Bayes (Expected A Posterior, EAP), and Bayes modal—we chose EAP because it has the smallest squared error integrated over the population distribution. Because we are now dealing with the probability sample represented by the total data, and not the groups assigned to the separate booklets, we may at this point reasonably assume the standard normal distribution for purposes of scoring all students.

Given the assumed population distribution and the estimates of the item parameters for the 3PL model, the EAP scores for the several content and process scales can be computed from subsets of the answer pattern of each student. Technically, the subset then determines the *posterior* (after-the-data) distribution for each student's proficiency on each scale. The mean of this distribution is the EAP estimate of proficiency for that subset, and its posterior standard deviation (PSD) is a measure of the estimator's precision, similar to a classical Standard Error of Measurement (SEM). The PSD or SEM is an indicator of the generalizability of the score in the sense that it conveys the range of variation that could be expected if another sample of items were used to estimate the same proficiency. The PSD or SEM varies, according to the level of the score, from one response pattern to another. This is one of the distinguishing properties of IRT scale scores compared with the number-right scores of classical test theory; the latter are assumed to have a constant SEM depending only on test length and the average item-trait correlation.

Once we compute the EAP scores for all students in the sample, we rescale them to have mean 0.0 and standard deviation 1.0 in the estimated distribution for the state. The rescaling makes use of the weights to make the sample "representative". In the present study, where all in-scope 8th-grade students were tested in each of the participating schools, these weights vary relatively little, and the weighted and unweighted statistics are very similar. The scale adjustments in the sample absorb the reduction in variance that occurs when EAP estimation is used, and they give scale score values that are almost identical to maximum likelihood scores, similarly scaled. Scores scaled to a specified mean and standard deviation in a representative (probability) sample are referred to as "standardized" or "normed".

Table 5.6 shows scale scores, standardized at mean zero and standard deviation one, for some students selected from Groups 1, 2 and 3. This scaling convention, which we refer to as the "0,1" standardization, is convenient for technical purposes. For public reporting purposes, this standardization is undesirable, because it includes negative values. In the reporting forms shown in Chapter 7, we therefore use the conventions of the California Assessment Program and NAEP—namely, to set the mean to 250 and the standard deviation to 50. We refer to this as the "250,50" standardization.

TABLE 5.6
Some typical scores and standard errors for students taking the Easy, Medium, and Difficult second-stage test (standard scores)

		Procedural Skill (n =15)			Algebra (n =9)		
	Student	Number Right	Score	S.E.	Number Right	Score	S.E.
Easy Booklet	SA	5	−1.28	0.55	1	−1.64	0.67
	NJM	8	−0.28	0.46	5	0.10	0.54
	TMO	13	0.932	0.50	7	0.75	0.55
Medium Booklet	PMC	3	−0.83	0.57	1	−0.94	0.60
	TJD	9	0.68	0.35	4	−0.23	0.59
	SRK	13	1.60	0.43	6	0.76	0.50
Difficult Booklet	RG	1	−0.83	0.65	3	−0.17	0.74
	RWK	9	1.59	0.36	4	0.94	0.60
	ASC	15	2.76	0.48	9	2.44	0.50

The standard errors of the scores in Table 5.6 can be used to estimate confidence intervals on the true proficiencies of the respective students. The interval bounded by plus or minus one standard error about a student's scale score has an approximately 2/3 chance of including the true value of the proficiency; the interval bounded by plus or minus two standard errors has an approximately 95 percent chance of including the true value. In displaying score profiles for the students, we represent one-SEM intervals by bars on each side of the estimated score (see Chapter 7). Notice in Table 5.6 that the standard errors are generally smaller for the students in Group 3; this reflects the location of the maximum of the test information above the mean for the California 8th-grade population.

Figure 5.4 shows the distribution of the 0,1-standardized scores for the total sample. Both distributions are unimodal and not excessively heavy in either tail. They tend, however, to be skewed toward the upper end of the scale. The skew is not so great as to vitiate standard statistical analyses of means based on normal distribution assumptions, given the robustness of these analyses to minor departures from normality. For most forms of secondary analysis, the distributional

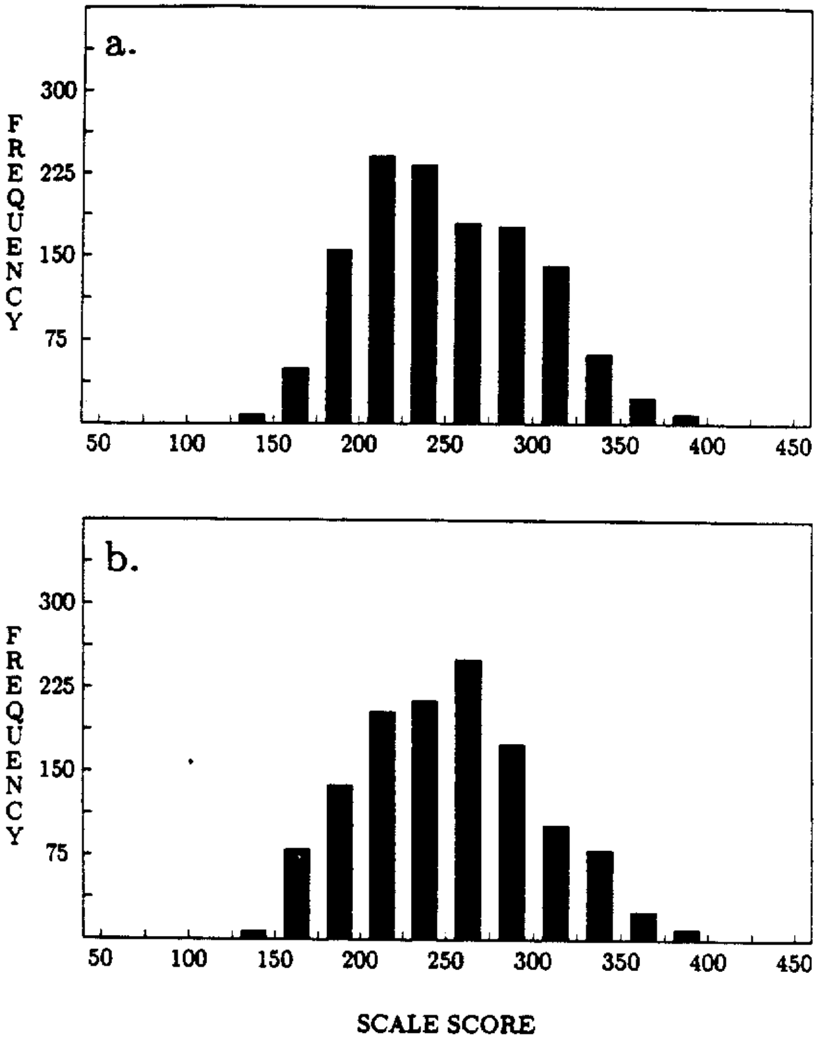


Figure 5.4. State-level distributions of the: (a) Procedural Skills scores, (b) Algebra scores.

properties of these two-stage student-level scores would be entirely satisfactory. But the skew *would* have to be taken into account in any models that attempt to predict percentiles of the state distribution. In Chapter 7 we offer a possible explanation for the tendency of scale-score distributions of mathematics attainment measures to be skewed toward the high values.

5.4 Information analysis of the second-stage test results

In IRT, the relationship between proficiency level and precision of measurement is expressed by the test information function. The concept of the information provided by a statistic derives from R. A. Fisher's theory of maximum likelihood estimation. But in the IRT context it corresponds to the reciprocal of the squared SEM at each point of the score continuum. Information is additive, so that if two subtests measuring the same trait are combined, the information for the total score will be the sum of both subtests' information functions. Similarly, the test information function is the sum of analogously defined information functions for the items that comprise the test.

Technically, test information applies to maximum likelihood estimates of scale scores. It is a function of the number of items, their slopes, and the value of the item response function at the scale point in question. For most tests, the information depends strongly on how the item locations are distributed along the continuum. If the items are spaced more or less equally from very low to very high values on the scale, the information function will be broad at the top, indicating that the test is accurate over a wide range of scores. But, more typically, the item will tend to cluster in the middle of the scale, and the information will be strongly peaked. In fact, it is rather difficult to assemble a truly broad-scale test with a flat information function. Even the pretest, which we attempted to make as broad as possible, has relatively concentrated information, as Figure 5.2 shows.

The mathematical expression for the test information function of maximum likelihood estimator of proficiency based on a logistic response model is as follows:

$$I(\theta) = \sum_{j=1}^{n} a_j^2 P_j(\theta)[1 - P_j(\theta)]$$

For EAP estimation, which we use in this report, the expression for *posterior* information is simply the above function plus a term accounting for the information contributed by prior knowledge of the population distribution of θ. For a normal population with standard deviation σ, the added term is $1/\sigma^2$, which in the case of an assumed unit normal distribution is just 1. If the number of items, n, is large, the contribution of prior knowledge is small relative to that contributed by the respondents' answers to the items and can be neglected. But with the short scales of the Duplex Design, the contribution is valuable and should be accounted for in the information function.

The properties of information functions are illustrated by the information plot for the pretest, shown in Figure 5.3. The figure includes the standard error curve, which is simply the square root of the reciprocal of the information function. Both curves show that the precision of the proficiency estimates varies as a function of proficiency level. The pretest provides considerably more information in the middle range of the ability continuum than at its extremes. This pattern is typical of most conventional tests. Because both relatively easy and difficult items contribute information in the middle range of proficiency, whereas difficult items provide very little information in the lower range and, easy items, very little in the upper range, information almost always peaks in the middle.

Calculation of the information for a two-stage test is more complicated than for a one-stage test because it varies not only as a function of the scale value, but also with the second-stage booklet the student is presented. Moreover, each booklet has its particular population distribution of proficiency, and the relative proportions of pupils completing each booklet is unequal. Before the information curves for the booklets can be combined to yield a curve for the form as a whole, each must be weighted by the population distribution of the relative proportion of pupils who completed the booklet at all points along the score continuum.

We performed these calculations for the Duplex instrument in four steps. First, we computed a posterior information curve and a normalized posterior distribution for each booklet of the form using the item-parameter estimates for the respective scale. Second, we derived weights for each booklet by multiplying the weights of the posterior distribution of ability for the subtest in that booklet by the population estimate of the proportion of pupils who completed that booklet. Third, we multiplied the posterior information curve of each booklet, as calculated at 81 points along the proficiency continuum, by the respective weights for that booklet. Finally, we summed the values obtained at each point to yield a posterior information curve for the form as a whole for that particular scale.

The information curves obtained in this way are shown for Procedural Skills and for Algebra in Figure 5.5. Because of our efforts to spread the difficulty levels of the second-stage booklets as widely as we could, the information curves are reasonably broad and have a gratifyingly flat region near the center. This means we are getting good measurement over a wider range than found in typical peaked achievement tests. But the problem of the difficulty levels of the forms is apparent in the location of the information peak in the 0 to 2 region of the population distribution rather than −1 to +1. This particular version of the instrument will perform better in estimating scale scores for better-performing students than for the poorly performing students. It would function well for all groups in Illinois, but for use in California, it should be revised to improve the information yield in the 0 to −2 region.

5.4.1 Reliability of the student-level scale scores

Figure 5.6 shows the corresponding SEM functions, computed as the square root of the reciprocal of the information functions. It is the squares of these error functions that are integrated over a (0,1) normal population distribution to obtain an estimate of the mean square error for the scales. This value and the estimate of the variance of the latent distribution of ability for the subtest in the form as a whole (as computed by BILOG using Sheppard's correction for grouped data) were used to obtain an estimate of the average reliability of the subtest

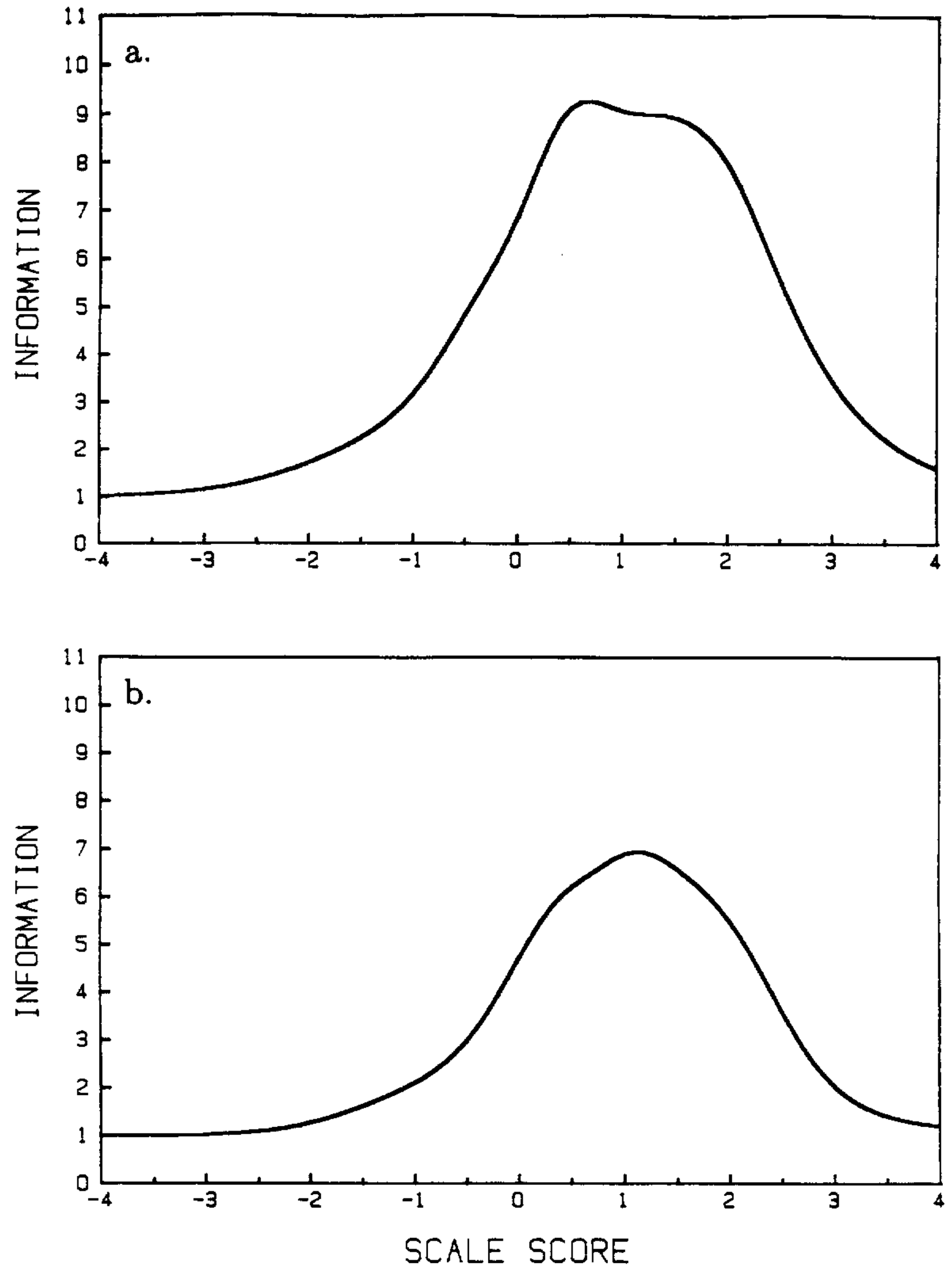


Figure 5.5. Information functions for the second-stage test scales: Form 1 (a) Procedural Skills, (b) Algebra.

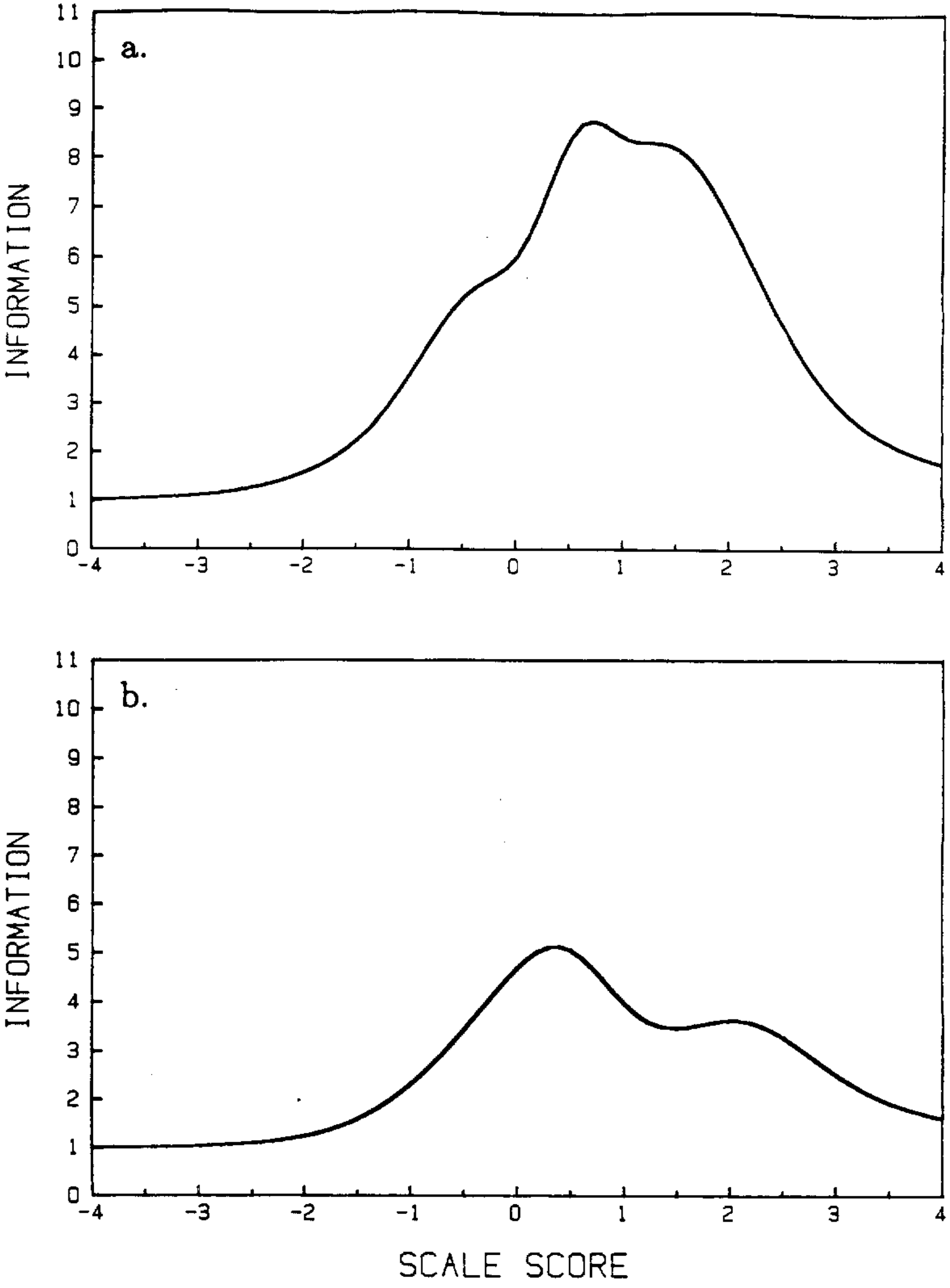


Figure 5.5 (continued). Information functions for the second-stage test scales: Form 3 (a) Procedural Skills, (b) Algebra.

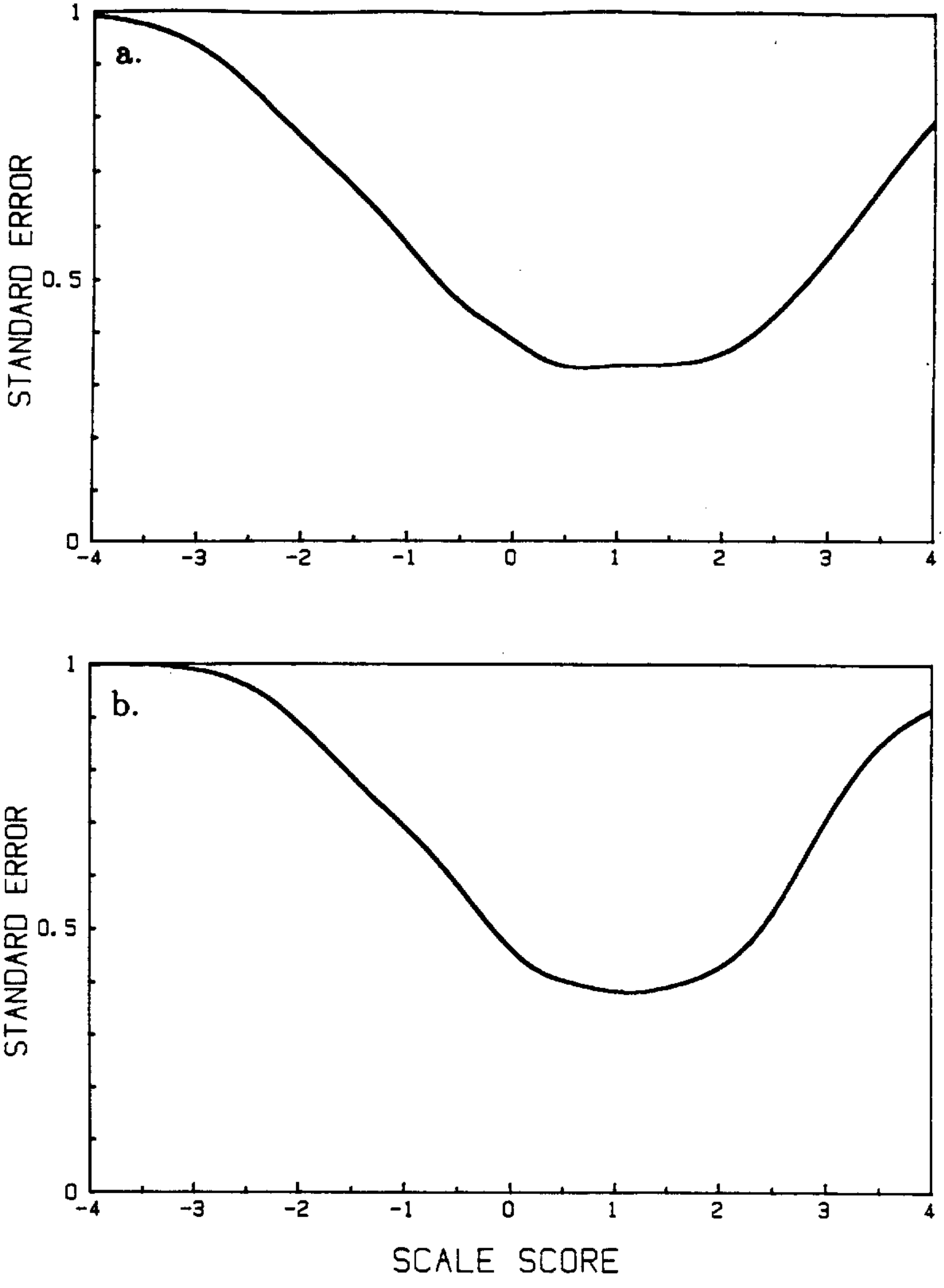


Figure 5.6. Standard error functions for the second-stage test scales: Form 1 (a) Procedural Skills, (b) Algebra

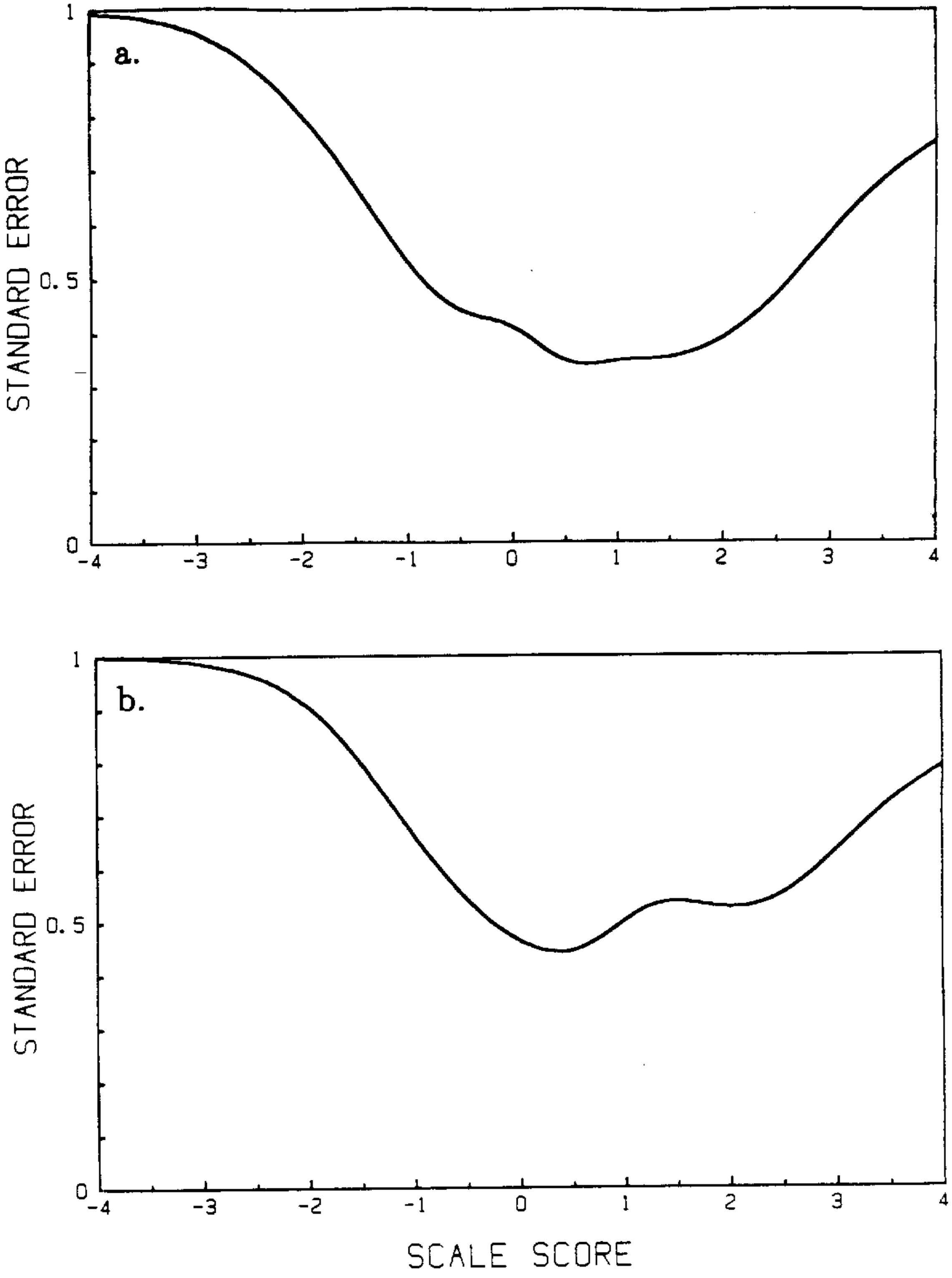


Figure 5.6 (continued). Standard error functions for the second-stage test scales: Form 3 (a) Procedural Skills, (b) Algebra.

TABLE 5.7
Reliabilities for two scales
in two forms of the second-stage test

Scale	Form 1	Form 2
Procedural Skills	0.871	0.841
Algebra	0.832	0.776

scores.

The overall reliabilities of the Procedural Skills and the Algebra scales of Forms 1 and 3 are shown in Table 5.7. The benefits of two-stage testing are evident in these values, which are gratifyingly high for these short scales. Although its scales are shorter, Algebra performs almost as well as Procedural Skills. This would be expected if the item composition of the content categories is more homogeneous than that of the process categories. Content is perhaps a more important source of variation in mathematics proficiency than is the type of cognitive process, as defined here. In addition, there is some variation in reliability due merely to the random assignment of items to forms.

All of these reliabilities are, however, high enough to justify the use of the scores for counseling individual students on relative attainments in main areas of content and skill. Moreover, when the five content scores are combined into an overall mathematics attainment score for each student, the reliability would be expected to increase to around 0.95, which is high enough to be used in decisions about the programs of individual students. Its slightly displaced difficulty level notwithstanding, the Duplex instrument of the California field trial meets or exceeds the reliability levels that would be needed in an operational assessment reporting at the student level as well as the school level.

But it is not just classical reliability that bears on the utility of the instrument: the broad information curve for the two-stage test also indicates better precision among students scoring toward the top and bottom of the distribution. These students, rather than the average students scoring near the center of the distributions, are the ones for whom critical placement decisions must most often be made. If the number of items is small, adaptive testing such as implemented in this study is the only way to assure accurate measurement at these extremes as well as in the middle of the attainment distribution.

5.4.2 Correlation between the scales, and the use of pretest information

The correlations between the Procedural Skills and the Algebra scales in the California sample in Forms 1 and 3 are .775 and .669. These fairly high values reflect in part the effect of the three items that these scales have in common. The separate content scales, or the separate process scales, share no items and should be less highly correlated. Even these correlations are well below the estimated reliabilities of the scales, however, so it is clear that these two scales are measuring more than one source of variation in mathematics attainment.

One consequence of the correlation between the scales is that the use of the pretest responses of the student to strengthen estimation of the scores in the content and process profiles would be ill-advised. Whatever it would add to the reliability of the separate scales would be lost in the resulting increased correlation between the scales. The score profiles would then contain very little information in addition to that contained in the student overall mathematics score.

But the pretest information would be useful in strengthening the overall mathematics score itself (which we obtain here by averaging the five content scores). Assuming the pretest items are a stratified random sample of the instrument, this can be done very simply, and optimally, by averaging the pretest scale score and the overall score from the second-stage, weighted inversely as their squared SEM's. In the present study, the reliability of the overall score would be high even without the pretest information. But in a design devoted, for example, to two-subject matter areas simultaneously, the number of items for each would be fewer and the area scores would benefit from the pretest information. In that situation, the pretest would presumably have half of its items drawn from each area, and each half would be scored separately before averaging with the corresponding second-stage score.

5.5 Gains due to two-stage testing

The test information functions of the previous section serve to evaluate the performance of two-stage tests as well as to guide their construction. When these curves are compared with a curve from a conventional test of equal length, each second-stage test booklet should exhibit gains in efficiency with respect to a one-stage test containing the same number of items drawn from all the booklets. For example, the Procedural Skills scale, based on 15 items in each of the three test booklets, could be compared with the performance of a 15-item, one-stage test composed of every third items from the three booklets. But, because the item information functions sum to form the test information, this would be same as dividing by three the overall information for the 45 items from the three booklets.

According to Lord & Novick (1968) (see also Lord, 1980), the relative efficiency function of, say, Test A to Test B is the value of the information function for A divided by that of B, at each point on the proficiency continuum. For purposes of illustrating the relative efficiency analyses for the Procedural Skills and Algebra scales in Forms 1 and 3, we carried out the calculation of the relative efficiencies in the following way. We began by computing a test information curve for the Easy, Medium, and Difficult booklets of each form and each scale. Then, we computed the full-test information curve for each scale by summing the curves for the three booklets over the score continuum. This curve shows the amount of information the test would provide if an examinee responded to the items in all three booklets of the form. Finally, we obtained the relative efficiency function of a given second-stage booklet as the ratio of the information function of that booklet to the full-test information function reduced by one-third.

The efficiency curves that result from these calculations are presented in Figure 5.7. They show the relative efficiency of each second-stage booklet compared with a one-stage test of the same length drawn from the same item pool. When both tests give equal amounts of information at a given level of proficiency, the height of the relative efficiency curve will equal one. When the second-stage booklet provides more information, the height of the curve will be greater than one. It shows how many times longer the one-stage test must be to

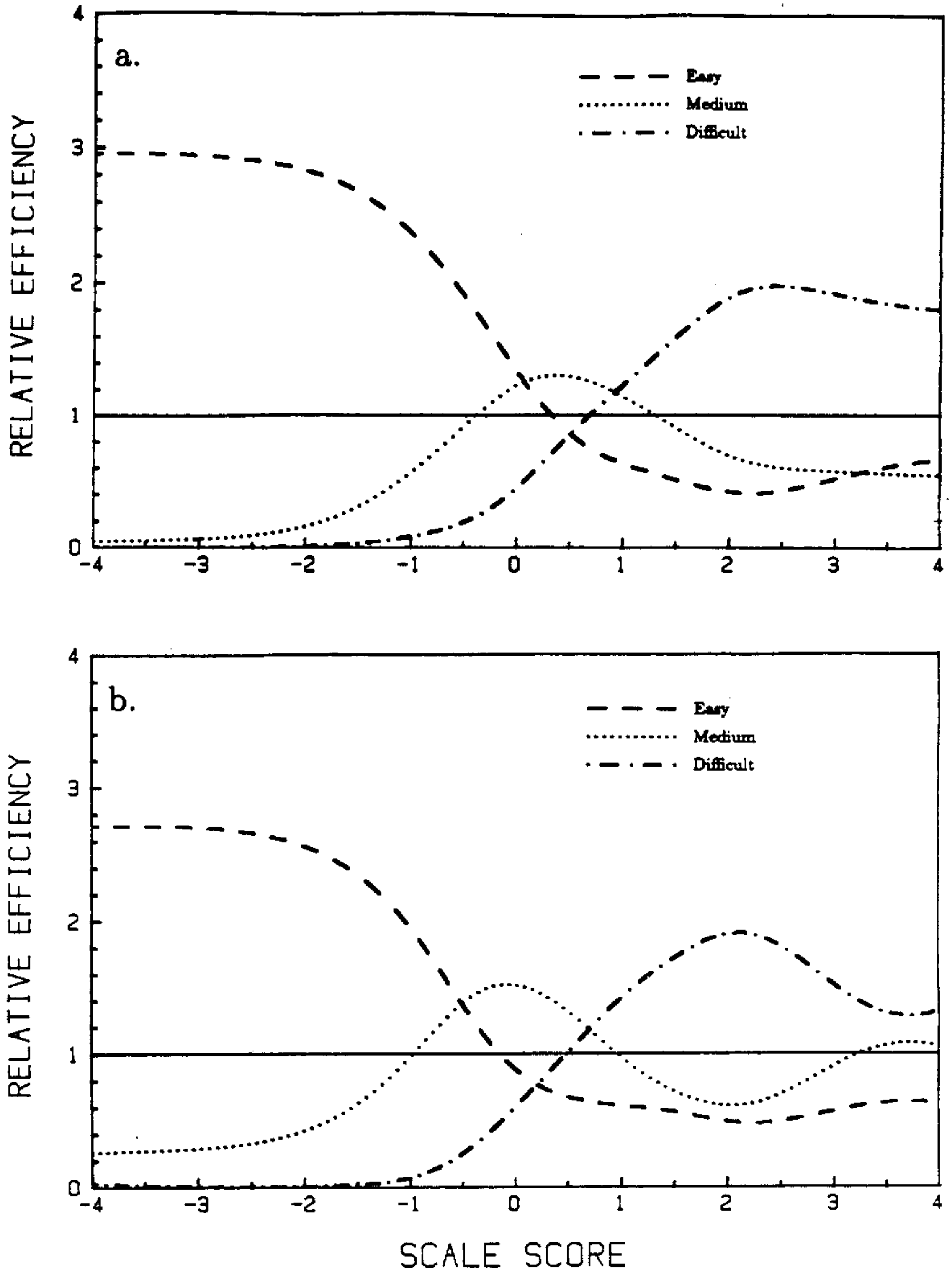


Figure 5.7. Relative efficiency of the second-stage test booklets: Form 1 (a) Procedural Skills, (b) Algebra.

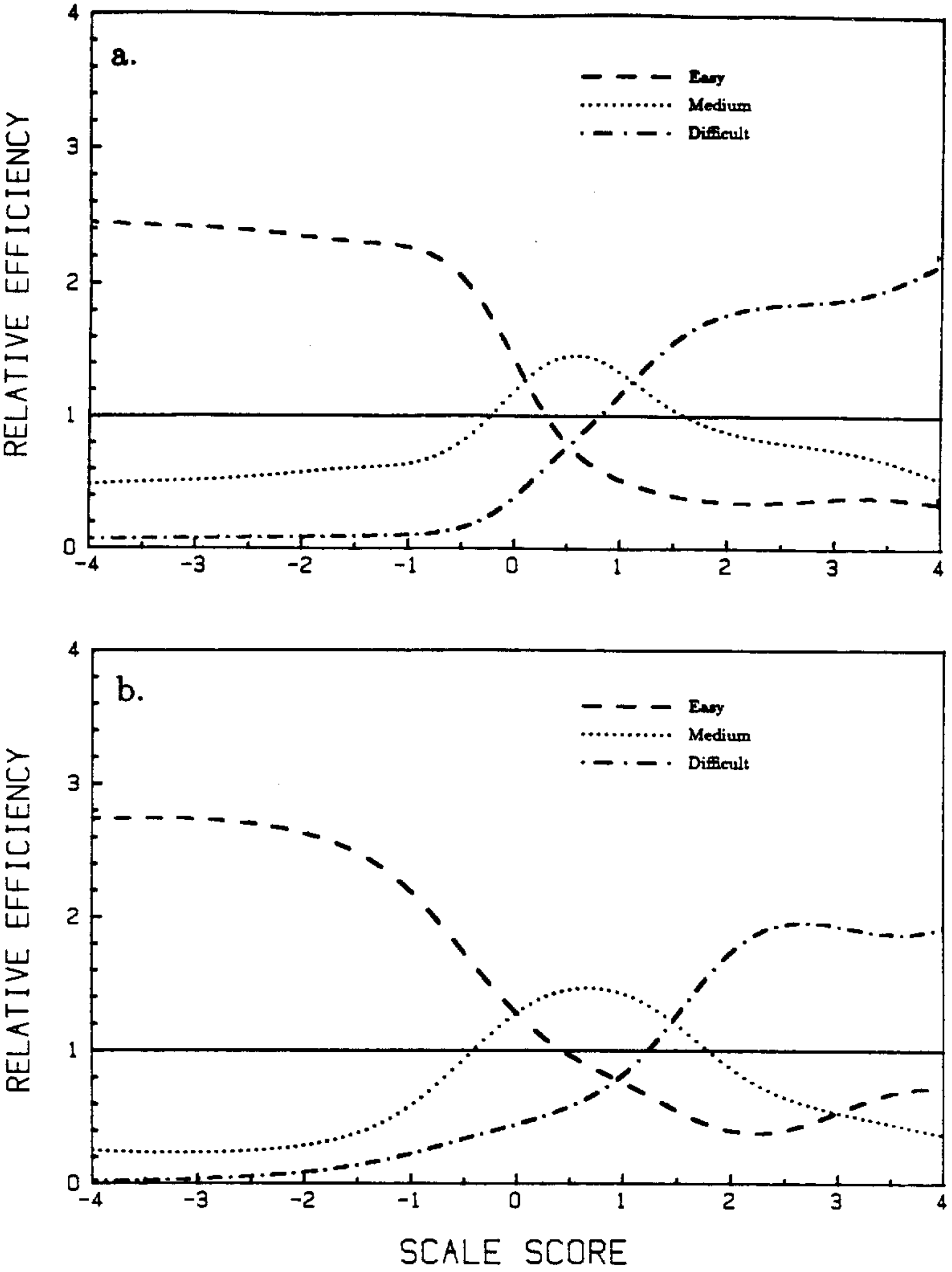


Figure 5.7 (continued). Relative efficiency of the second-stage test booklets: Form 3 (a) Procedural Skills, (b) Algebra.

provide the same amount of information as the two-stage test at that level. When the booklet exhibits a loss in efficiency, the height of the curve will be less than one; *i.e.*, the one-stage test is more informative than the second-stage test. It is apparent in Figure 5.8 that, in the range of scores to which they apply, the Easy and Difficult booklets are two to three times more informative than a one-stage test. Thus, to obtain the same measurement precision in these ranges, a one-stage test would have to be two to three times longer than the two-stage test. In the middle range, the gain is smaller because, as for most tests, the item locations for this test peak toward the middle of the scale. Even here, however, the two-stage test is comparable to a somewhat longer one-stage test.

These relationships can perhaps be seen more clearly in the relative efficiency curve for the second-stage test as a whole, shown in Figure 5.8. To obtain this type of efficiency curve, we weighted the relative efficiency curve from each booklet by the posterior distribution of ability found with that booklet in the California sample. After each of these curves was weighted by the proportion of pupils who responded to the respective booklet, they were summed to yield a relative efficiency curve for the form as a whole in each area of the content-by-process design. These curves show the average gains in efficiency found with the instrument, given the proportions of students assigned to the second-stage booklets.[4]

The efficiency curves again show that the largest gains of two-stage over one-stage testing occur toward the extremes of the distribution, where conventional tests often have limited accuracy. Toward the center, the one-stage test is difficult to improve upon, although some gain is evident. The curves for the two forms are similar, as expected, and could be combined with information curves averaged across forms for purposes of displaying the overall efficiency of the instrument.

[4] The irregularities at the extremes of these curves are due to instability of the calculations with extremely small probability values.

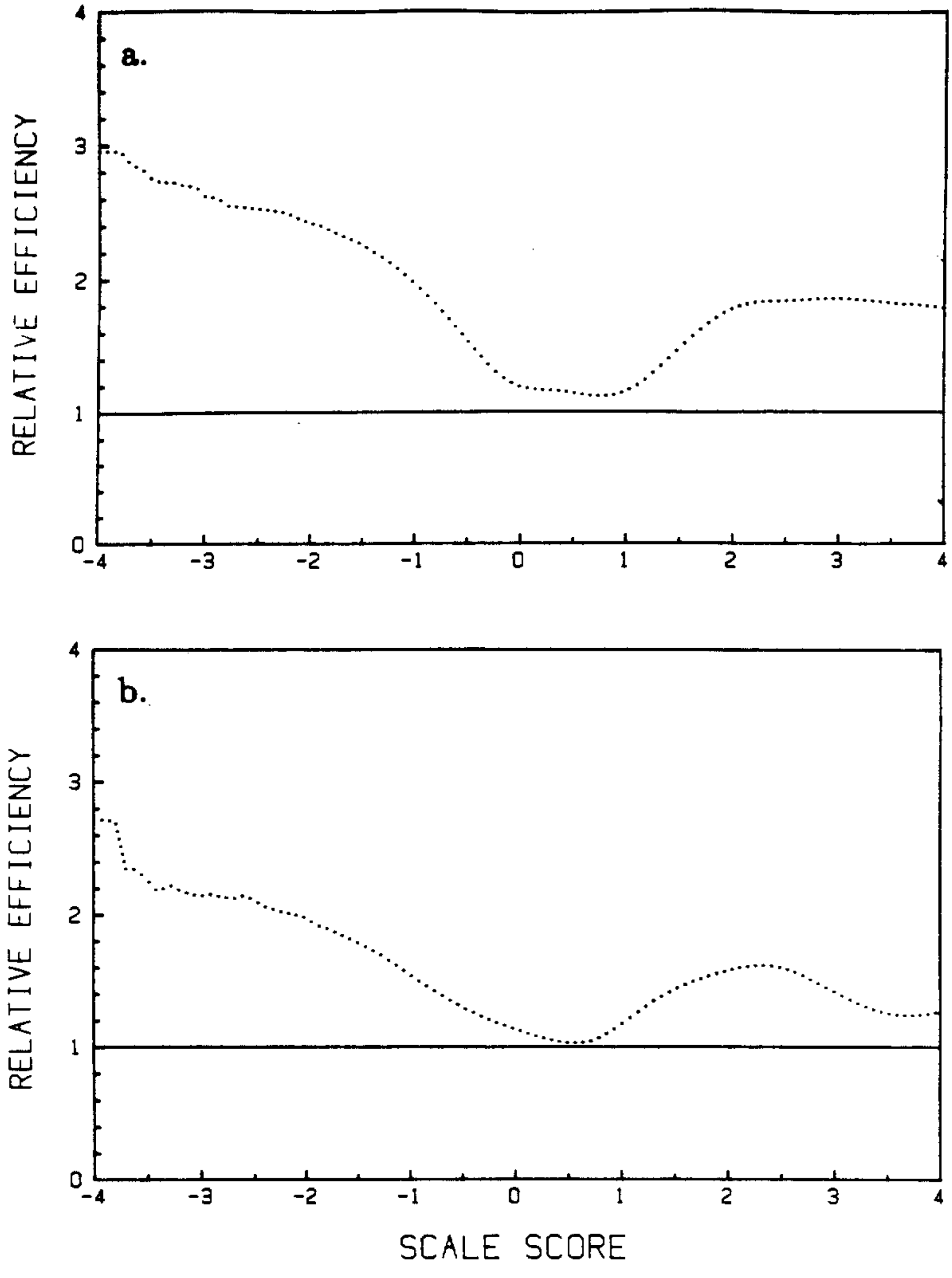


Figure 5.8. Relative efficiencies of the second-stage forms as a whole: Form 1 (a) Procedural Skills, (b) Algebra.

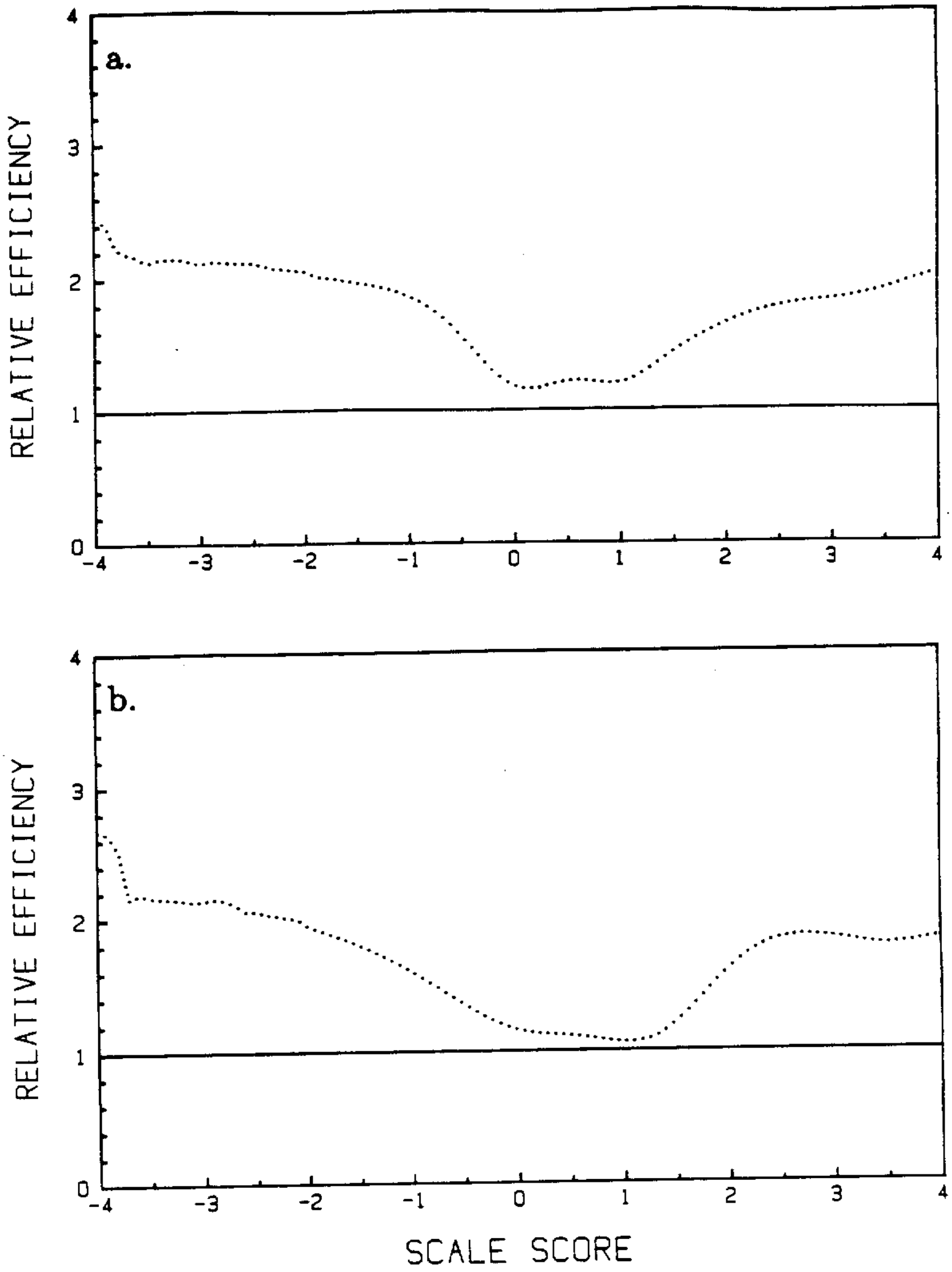


Figure 5.8 (continued). Relative efficiencies of the second-stage forms as a whole: Form 3 (a) Procedural Skills, (b) Algebra.

5.6 Summary

The questions we raised at the beginning of this chapter on student-level scoring can now be answered, largely in the affirmative.

1. The procedure we adopted for administration of the two-stage assessment instrument appeared to have good teacher acceptance, and it posed no problems for the students. Administration of the first-stage test one or two days before the second-stage test had a number of advantages. It gave the classroom teacher time to explain the nature and purpose of the testing more adequately; it allowed more time for the students to become familiar with the answer sheet by entering their names and background information. Most importantly, the 15-item pretest gave the students practice on the type of item that would appear in the main test. As for the teachers, they or their assistants found the transparent scoring-template for the pretest easy to use, and in general they followed accurately the instructions for assigning students to the second-stage test booklet according to pretest scores. The field trials revealed no insurmountable difficulties in implementing teacher-administered two-stage testing by this method in large-scale assessment programs.

2. We were gratified to find that our 8th-grade mathematics items—drawn from the Illinois and California assessment programs, the Second International Mathematics Study, and other sources—had generally high discriminating powers. Because of the quality of the items and by use of highly efficient methods for scoring the item responses (two-stage testing, differential weighting of responses according to item discriminating power, Bayes estimation combining prior and posterior information about student proficiency), we attained reliabilities for the content and process categories generally in the 0.80's and, for the overall mathematics score, a predicted reliability of 0.95. We consider the reliabilities for the student-performance profiles quite adequate for student and parent counseling, and those for overall mathematics attainment adequate for student placement and certification.

3. The posterior standard deviations (essentially, standard errors of measurement) of the student profile scores, even from the rather short content scale, were less than 0.5 sigma units (S.D.'s) over a range of three standard deviations of the population distribution. The distribution of scores in the full 5,023-case sample was unimodal with well-behaved tails, and skewed somewhat toward the high end, as we expect for a state-wide distribution of mathematics attainment at this grade level. The scores would perform well in any form of statistical analysis that assumes a normal distribution for group means, as well as in multilevel analyses that compute within- and between-school variance components (see Chapter 6). Inasmuch as the scoring procedure yields the posterior mean and standard deviation for each student, they supply the first-level sufficient statistics for marginal maximum likelihood or empirical Bayes estimation in multilevel analysis (see Bock, 1989). These properties of the student-level scores from the two-stage Duplex Design are well-suited to secondary studies of assessment data aimed at separating and elucidating student effects and school effects.

4. Our analyses of the relative efficiency of the two-stage scales, compared to a one-stage scale with the same number of items, showed the former to be more than two times as efficient as the latter in the upper and lower thirds of the score distribution where critical placement and certification decisions are most often made. In the less critical middle third, the two-stage scales still showed some gain in efficiency. We obtain these gains even though our second-stage booklets, which we assembled from item statistics from the Illinois study, were somewhat too difficult for the California population. Moreover, we used more link items between booklets than ultimately proved necessary, thus reducing somewhat the effectiveness of the two-stage instrument. That the two-stage procedure worked well even in these adverse conditions leads us believe that substantial gains in the quality of educational measurement by these methods can be obtained rather easily. It suggests that any large-scale testing program, whether assessment or traditional achievement

testing, that is now using one-stage testing is missing an opportunity to reduce testing time by at least one-half.

Considering the competing demands of external testing programs for classroom time, we believe that the power of two-stage procedures to reduce time for student-level testing by one-half or more abundantly justifies the additional effort involved. Once the test administrators become familiar with the method, and if they have some help in scoring the pretest, two-stage testing could become a standard feature of large-scale assessment programs. In our trials of the procedure, both in Illinois and California, nearly all aspects of instrument construction, administration, analysis, and scoring of the two-stage Duplex Design proceeded as expected and, in its operating characteristics, performed even better than expected.

The analysis and scoring of the data from the two-stage Duplex Design reported in this chapter are meant only to demonstrate how the procedures can be carried out and what results can be expected. All of these steps described can be performed on any marginal maximum likelihood item analysis and Bayes scoring program. (We used the BILOG program of Mislevy and Bock in the 1983 mainframe version and also the 1987 PC version.) In an operational state or national assessment using these methods, a larger system capable of constrained estimation over the randomly replicated forms would be needed to obtain optimal results while minimizing the need for intervention by the data analyst. The present study provides empirically tested guidelines for the development of such a system.

Chapter 6

Measurement at the school level

The original conception of matrix-sampling designs in educational measurement, as introduced by Lord in 1962, was to establish test norms with minimum intrusion on classroom time by drawing a random sample of students from some population and administering to each a small random sample of items from some domain of content. The overall percent of correct responses in the total data would then provide a highly generalizable index of average domain mastery of students in the population. Moreover, the cost of data collection would be reduced because the economies of statistical sampling would apply to both the students and the test items. This was the the new approach to measurement for large-scale assessment that NAEP pioneered in the 1960's and 70's. Using a low rate of sampling students, NAEP was able to report average percents correct in several subject matters at three grade levels in four regions of the United States. In this way, it could monitor general progress in education at the national level and compare performance in the regions.

As the states developed their own assessment programs, however, many chose to test *all* students at selected grade levels and to confine the sampling to the item domains. These so-called *census* assessments made possible the reporting of average percent correct, not only for the state as a whole, but for units as small as individual schools. Using multiple matrix-sampled test forms, a census assessment could gather a wealth of statistics on progress of individual schools toward detailed objectives of the curriculum without requiring more than one class period of testing time per student. In this type of analysis, the school is the primary unit of analysis and reporting. Technically, groups as small as classrooms could be primary units. Because of the possible confounding of instructional effects with those of ability-grouping by classrooms, census assessment results have seldom been reported below the school level.

The adoption of schools as the unit of analysis has the advantage of facilitating the use of IRT scaling, rather than averaging of item percents correct, as a method of scoring the assessment instrument. Scaling provides a way of maintaining comparability of the reported scores as the item content is updated from time to time. Bock, Mislevy & Woodson (1981) and Mislevy (1983) have shown how an IRT model can be formulated that allows a score measuring average performance in the school to be computed directly from the item responses of the students. With a suitably designed instrument, this group-level IRT model makes the analysis and scoring of the matrix-sampling data straightforward and computationally efficient. For each school, it produces scale-scores in each of the specified curricular objectives, and these scores can be summarized for skills, topics, and subject matter, as well as aggregated at the district, county, and state level. In California, these methods have been employed successfully since 1980 for reporting the assessment results for grades 3, 6, 8, and 12.

In view of the trend toward census assessments in most states, we have assumed school-level measurement of curricular objectives in the present studies of the Duplex Design. Except for the complications introduced by two-stage testing, we employ the same methods in estimating scale scores for the schools as those developed for the California Assessment Program (see Bock & Mislevy, 1981). The schools are assumed to be random units of observation, playing the same role in the group-level IRT analysis as the students play in the individual analysis. The scale values for the schools, which represent the average attainment of the respective students, are assumed to have a distribution on an underlying dimension for each of the objectives assessed. The item parameters of the school-level response models are estimated by the marginal maximum likelihood method just as in the student-level estimation in Chapter 5. In these analyses, the original item response records, which already have been used twice in computing the within-booklet content and process scores for individual students, are now combined across booklets and forms to estimate the curricular-objective scores for the schools.

Although these group-level methods treat the schools as the primary unit of analysis, the summary scores for the distinct, county, and state are obtained by weighting school scores by the number of students at the respective grade levels. This applies to the standard deviation of school scores as well as the mean. Because the standard errors of the school scores tend to be proportional to the numbers of students in the schools, these weighted statistics are essentially minimum variance estimators of the state mean and standard deviation.

Such aggregate statistics are quite suitable for monitoring general progress of the state educational system and for examining school effects, but they contain no information about student or classroom variation within-schools, and thus cannot be used in multilevel data analysis. The student level scores described in Chapter 5 serve that purpose, though necessarily with less detail than is required for the evaluation of school-level curricular objectives. It is for this reason that both student-level and school-level scoring is required in a comprehensive assessment program.

6.1 Assumptions of school-level measurement

In order to meet the assumption of conditional independence of item responses in IRT scaling at the school level, each item assigned to a given objective must appear in a separate test booklet, and thus represent the response of a different student. Assuming that the students do not collaborate, the responses for any given scale will be statistically independent within each school. In the present study, where the Duplex Design consists of 24 test booklets each with 45 items, scales can be constructed measuring the 45 distinct curricular objectives defined in Table 4.1. For reasonable accuracy of the school-level scores, it is preferable for the school to have at least 24 students at a grade level, so that each distinct booklet is used at least once. If there are very small schools in the system, as is true in rural California, special Bayesian methods must be employed to obtain stable scores at the school level. We will not discuss these details here; they are available in technical reports of the California Assessment Program.

The sufficient statistics for estimating school-level scores take the compact form shown in Table 6.1. For each school, the data consist of the number of students presented each item of each test booklet, and, among those, the number who responded correctly. Except for the link items, students in Groups 1, 2, and 3 are responding to different items. We refer to this table as the "number-tried, number-right" school summary. For small schools, such as number 1 in Table 6.1, the incidence of item presentations will be rather low and will be reflected in larger standard errors for the school score. For very large schools, such as number 3, the school score for each objective will be much more precise.

Under the assumed sampling conditions, the number of right responses, given the number of items presented, is an independent binomial variable for which statistical modeling is entirely straightforward. To extend the IRT models to the group level it is only necessary to substitute for the product-Poisson probability of an answer pattern given by equation (5.2) in Section 5.2, the product-binomial probability. The result is as follows:

$$P(\tau, N) = \prod_{j=1}^{n} \frac{N_j!}{r_j!(N_j - r_j)!}[P_j(\tau)]^{r_j}[1 - P_j(\tau)]^{N_j - r_j} \qquad (6.1)$$

In (6.1), τ is the proficiency level of the school for the objective being scored, $P_j(\tau)$ is the item response function, n is the number of items, and N_j and r_j are, respectively, the number tried and number right for item j. The BILOG program (Mislevy & Bock, 1983) has provisions for estimating the parameters of the item response function and for scoring the group with this model when the data consist of number-tried, number-right summaries.

The use of a group-level model can be justified as a parsimonious way of accounting for data in the form of Table 6.1. If the data conform to the 3PL model, for example, the entries in a table such as 6.1 can be accurately predicted from a scale score for each school and three parameter values for each item. Experience with California assessment data has consistently shown that the group-level logistic models reproduce this type of data very well. Mislevy (1983) has proposed a threshold process that accounts for the highly satisfactory performance of these models. In the Duplex Design, their goodness-of-fit benefits from the highly homogeneous item content in the narrow curricular objectives for which scales are constructed. The assumption of unidimensionality is easier to satisfy in this context than in the more heterogeneous content and process categories of student-level measurement.

Unlike the right-wrong data at the individual level, the number-tried, number-right data are amenable to a straightforward test of the goodness-of-fit. The fact that different students respond to each of the items of any given scale justifies a conventional chi-square test of independent binomial variables: the data supply the observed frequencies, and the number-tried times the value of the 3PL model supplies the expected frequencies for the particular school. There is one such chi-square value for each school, and the values may be summed over the schools to obtain an overall test of the goodness of fit. The number of degrees of freedom is the number of independent entries in the data table minus the number of school scores and item parameters estimated.

6.2 Estimating item parameters of the school-level model

Except for the substitution of product-binomial probabilities for product-Poisson probabilities, the fitting of the school-level model is similar to that of conventional individual-level IRT models. In place of a proficiency value for each student, there is an average proficiency value for each school. The schools are assumed to be a sample, or a total census, of a population of schools, and the corresponding average proficiency values are assumed to have a distribution in that population. As in the individual-level analysis, this distribution is integrated over θ to obtain the marginal probabilities of the patterns of number-tried, number-right data in the sample. The MML method is then applied to assign estimates to the items parameters so as to maximize the marginal probability. Because the amount of information is greater in group-level data than in individual level data, relative to the number of parameters fitted, estimation for the group-level model

TABLE 6.1
An example of two-stage number-tried (T) and number-right (R) data
for an objective of the Duplex Design

		Item															
School	Group	1		2		3		4		5		6		7		8	
		T	R	T	R	T	R	T	R	T	R	T	R	T	R	T	R
1	1	5	3	4	1	3	2	3	1	2	0	4	3	4	3	5	3
1	2	5	2	3	2	4	2	2	2	3	2	4	2	5	2	4	3
1	3	2	2	2	2	5	4	3	3	4	3	2	2	3	2	2	1
2	1	10	4	9	3	9	5	8	3	8	4	7	2	6	1	8	5
2	2	6	3	5	2	4	2	7	3	7	3	8	4	8	4	9	3
2	3	5	4	6	4	6	6	6	3	4	3	6	5	6	6	6	6
3	1	34	22	33	16	32	21	36	19	36	12	35	21	34	21	34	15
3	2	23	18	24	17	24	15	22	13	22	7	23	11	23	6	20	13
3	3	26	13	20	9	24	11	21	9	18	12	22	12	19	4	20	7
etc.																	

is highly stable and the standard errors are very small. In addition, the assumption of a normal population distribution is easier to justify for schools because the scores for average performance of students benefit from the tendency of means of independent observations to be normally distributed.

When the school-level data derive from two-stage testing, number-tried, number-right frequencies for students are assigned to each type of second-stage booklet in each school. Table 6.1 is an example of such data for one of the objectives. Note that, typical of matrix-sampling designs, the data extend across forms. As is also the case within forms, different items appear in the Easy, Medium and Difficult booklets of the eight replicate forms, apart from the link items. When the data are scored, scale scores must be computed separately for the three second-stage groups. Then the average of these three scores, weighted by the numbers of students in the groups, are computed to obtain the scale score for the school.

Similarly, the item parameters are estimated separately in the three groups, and the link items are then used to express the estimates on the same scale. As in the individual-level item analysis in Chapter 5, there is again the problem in MML estimation of the provisionally assumed population distribution for each of the second-stage groups; the sample size in the present study is not large enough to estimate the separate population distributions concurrent with the item parameter estimates. Our solution to this problem is the same as in Chapter 5; we will initially infer the latent distributions for each of the groups from the pretest data; then, while restricting the slope and guessing parameters of the link items in the three groups to be equal, we will adjust all of the location-parameter estimates so that the differences of the link-item locations sum to zero.

It may seem strange to use the pretest latent distribution for individual students as the provisional population distribution for the schools, but in fact the latent distributions of the group means will have the same weighted mean as the student-level distributions. The standard deviation will differ, but that is absorbed in the slope parameters of the group-level model. The units of the resulting scale will, of course, be entirely arbitrary, but we will adjust them so that the scale of the school-level scores will be consistent with that of the student-level scores. Our method of making this adjustment is described in the next section.

To illustrate group-level item analysis, we show the results for the Fractions Concepts objective of the Numbers content area. This objective is represented by 20 distinct items, four of which are link items. Provisionally, we express the item-parameter estimates on a scale set to mean zero and standard deviation one in the latent distribution of *schools*. The estimates appear in this form in Table 6.2. Compared to the student-level estimates in Table 5.5, the values of which are expressed in the 0,1 metric for individuals, those in Table 6.2 vary more in their locations and have generally smaller slopes. Otherwise, there is nothing to distinguish them from the results of an item analysis based on an individual-level model.

6.3 Expressing the student-level and school-level scores on the same scale

In an exclusively school-level assessment, there is no plausible alternative to expressing the scores on a scale standardized in the sample of schools in the state. The scales of the California Assessment Program, for example, are standardized in the base year so that the weighted mean and standard deviation are 250 and 50, respectively. With the Duplex Design, where scale scores are available both for students and for schools, we have the possibility of standardizing at either level. For consistency with the conventions of traditional educational measurement, the natural choice is to standardize in the student-level data and then express the school scores in those units. Mislevy & Bock (1989) have proposed a type of IRT analysis that obtains the student and school scores jointly on the same scale, but their method has not been adapted to two-stage testing. We have chosen instead to set the scale of the school scores so that they have the same variability as the school means computed from the student level scores. Although this can be done in a ıumber of different ways, we have elected to set the unit of the school-level scores so that the residuals from the model used in predicting the school scores from their background characteristics have the same variance as similar residuals computed from the

TABLE 6.2
School-level item-parameter estimates
for Fractions—Conceptual Understanding

Form	Item	Slope	Location	Asymptote
1	5E	1.243	−0.979	0.173
2	5EM	0.525	−0.892	0.182
3	5E	0.401	−1.761	0.211
4	5EM	0.798	0.101	0.197
5	5E	0.534	−2.753	0.186
6	5E	0.616	−0.535	0.211
7	5E	0.697	−0.817	0.203
8	5E	1.311	0.171	0.156
1	5M	0.698	1.159	0.231
2	5ME	0.525	−0.892	0.182
3	5MD	1.048	0.796	0.170
4	5ME	0.798	0.101	0.197
5	5MD	0.570	−0.141	0.192
6	5M	0.886	0.833	0.215
7	5M	0.584	0.448	0.203
8	5M	0.938	0.421	0.199
1	5D	1.262	1.241	0.162
2	5D	0.884	1.898	0.182
3	5DM	1.047	0.796	0.170
4	5D	0.702	−0.091	0.169
5	5DM	0.570	−0.141	0.192
6	5D	0.600	1.827	0.185
7	5D	0.576	1.614	0.187
8	5D	1.478	1.607	0.168

TABLE 6.2 (continued)
School-level item-parameter estimates
for Equations—Procedural Skills

Form	Item	Slope	Location	Asymptote
1	16E	0.485	−2.329	0.195
2	16E	0.486	−1.608	0.208
3	40E	2.830	−0.192	0.500
4	40EM	0.794	0.834	0.185
5	31E	0.656	−0.403	0.223
6	31E	0.612	−1.100	0.200
7	22E	0.655	−0.413	0.213
8	22EM	0.499	−1.408	0.177
1	16MD	0.812	0.715	0.168
2	16M	0.940	0.932	0.157
3	40MD	0.636	0.595	0.193
4	40ME	0.793	0.834	0.185
5	31MD	0.883	0.939	0.156
6	31MD	0.828	3.321	0.103
7	22M	0.544	−0.312	0.185
8	22ME	0.499	−1.408	0.177
1	16DM	0.812	0.715	0.168
2	16D	0.785	0.476	0.167
3	40D	0.505	3.872	0.168
4	40D	0.593	3.645	0.143
5	31DM	0.884	0.939	0.156
6	31DM	0.828	3.321	0.102
7	22D	0.771	1.881	0.157
8	22D	0.679	2.491	0.185

TABLE 6.2 (continued)
School-level item-parameter estimates
for Figures—Problem Solving

Form	Item	Slope	Location	Asymptote
1	24E	0.420	−0.506	0.208
2	24E	0.641	−0.069	0.196
3	15E	0.508	−0.297	0.179
4	15EM	0.285	−1.337	0.215
5	39E	0.837	1.298	0.144
6	39EM	0.678	0.123	0.216
7	30E	0.639	0.097	0.181
8	30E	0.740	0.082	0.182
1	24M	0.615	1.394	0.216
2	24M	0.500	0.461	0.197
3	15ME	0.508	−0.297	0.179
4	15MD	0.509	0.186	0.196
5	39ME	0.836	1.298	0.144
6	39M	0.590	0.358	0.186
7	30ME	0.639	0.097	0.181
8	30MD	0.461	1.571	0.235
1	24D	0.526	1.480	0.184
2	24D	0.827	2.157	0.128
3	15D	0.482	1.558	0.193
4	15DM	0.509	0.186	0.196
5	39D	0.532	1.915	0.168
6	39D	0.549	2.239	0.157
7	30D	0.599	1.099	0.169
8	30DM	0.461	1.571	0.235

TABLE 6.3
Mean squares from the analysis of
variance of school-level scores and
school means of student-level scores for
the Algebra Equations, Procedural Skills objective

		Mean Squares	
Effects	df	School Scores	School Means
Main class	5	102.222	58,026
2-factor interactions	10	31.015	18,491
Higher-order interactions[1]	12	25.051	23,652

[1] Four degrees of freedom are lost because of empty cells.

school means of the student scores. Because that model is the basis of certain forms of displaying assessment data described in Chapter 7 (especially Figure 7.2), this method of scaling makes these displays consistent whether the school-level scores or the means of student-level scores are depicted.

This method of setting the unit of the school-level scores depends upon analyses of variance of the 2^5 sampling-allocation design described in Chapter 4. One such analysis is performed on the school means of the student-level scores as expressed in the 250,50 standardization of Chapter 5. The other is performed on the school-level scores estimated from the item parameters of Table 6.2, which are in the arbitrary 0,1 metric in the latent distribution of schools. In both analyses, we assume that all main effects and two-factor interactions of the five-way design are fixed, and that the remaining three-, four-, and five-factor interactions are random. For the Algebra Equations, Procedural Skills objective, for example, the mean squares for effects in the two analyses are shown in Table 6.3.

For the multiplicative scaling constant that converts the school-level scores from the 0,1 metric to the student-level 250,50 standardization we use the ratio of the root mean squares of the respective higher-order interactions:

$$\begin{aligned} r &= 153.7921/5.0051 \\ &= 30.7271 \end{aligned}$$

After the school-level scores are multiplied by this constant, their weighted mean is set to 250 to agree with that of the student-level scores. It is on this scale that the school profiles for curricular objectives appear in the reporting forms illustrated in Chapter 7.

6.4 The estimated state distribution of school-level scores

In the standardization derived in the previous section, the estimated distribution for the Equations-Procedural Skills objective of the schools in California that include grade 8 is represented in Figure 6.1. Compared to the distribution of student-level scores for the Algebra content area, the standard deviation is much smaller and, because this is in effect a distribution of means, the skew seen in the student score distribution (Figure 5.4) is no longer evident; the distribution appears essentially normal.

6.5 Standard errors of combined scores

To obtain overall mathematics scores at higher levels of aggregations, such as the district and state, one has the option of averaging the student-level scores over the content categories, or averaging the school-level scores over objectives. We prefer the latter, on grounds that the scaling of the school-level models meets its assumption with somewhat more plausibility than the student-level model. The overall mean for all objectives, or main area of attainment, is averaged over schools, weighting by the numbers of students in the schools.

Estimating the standard errors of these mean subject-matter, skill, or content area scores would seem to present a problem, however, for they are based on the responses of the same students, and the measurement errors are necessarily correlated. We have estimated the error correlations by a method we call the *split-school* technique.

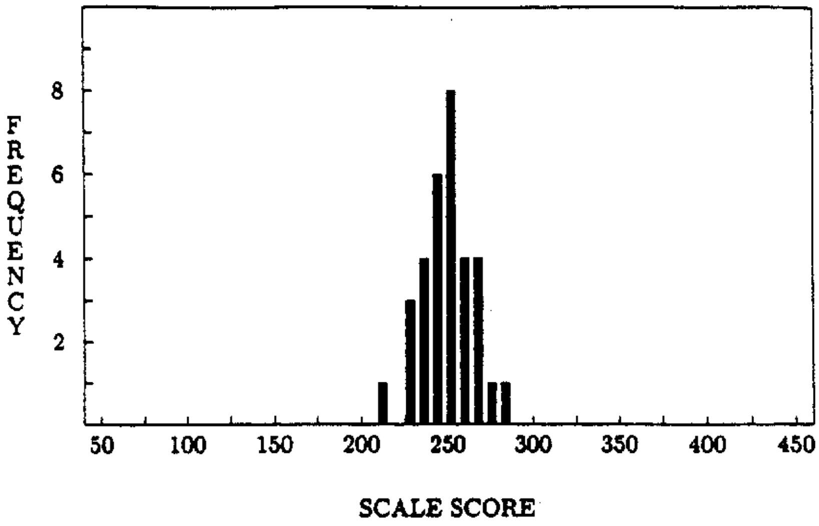


Figure 6.1: Estimated state-level distribution of school-level scores for the Equations-Procedural Skills objective.

Excluding very small schools in the sample, we randomly split the data of each school into two groups. Then we compute the scale scores for the split schools in the usual way. From these paired scores, we compute the pairwise product-moment correlations among all the objectives. For those objectives for which an overall or area mean score is required, the conventional formula for the standard error of the mean of correlated variables applies (see Lord & Novick, 1968, Chapter 4). The split-school technique assumes, of course, that the correlation structure of the errors is homogeneous among schools.

6.6 Summary

The extension of IRT models to permit the scaling of average performance of groups of respondents directly from item responses, without calculation of scores for individual respondents, is a productive development in assessment methodology. In addition to simplifying the maintenance of consistent scales of measurement as items are retired and replaced in the assessment instrument, it makes the efficiencies of two-stage testing available to matrix-sampling designs. Many more items in the instrument can be selected for an approximately 0.5 difficulty level in the second-stage groups, thus increasing substantially the information in the item responses. These gains in efficiency translate into more accurate estimation of curricular objectives, or of greater detail in the objectives assessed.

This chapter demonstrates how item-parameter estimation and scoring of two-stage, group-level data parallels that of the analysis of individual-level data described in Chapter 5. It is only necessary to assume a population of groups, which in the context of large-scale educational assessment corresponds to the schools in the state or nation, and proceed with group-level item response models fitting by the marginal maximum likelihood method. The fitted models serve to predict, as a function of a scale value for the school, the proportion of students in each school who will respond correctly to each of the items in the matrix sample.

The scale values, or scores, for the schools then bear the same relationship to the student-level scores as group means bear to the observations within-groups. By methods described in this chapter, the IRT scale scores for schools can be expressed in the same units as the student-level scores of the previous chapter. The result for assessment is a unified system of measurement suitable both for reporting purposes and for secondary studies of student and school effects in instruction and learning.

Chapter 7

Reporting[1]

The Duplex Design provides for reporting assessment results on a number of levels—for the state as a whole, for counties and districts, for schools, classrooms within schools, and for the individual students.[2] The reporting forms will vary according to the amount of detail needed at each of these levels, and they will be more effective if they display the results graphically as well as numerically. In this chapter we illustrate how the reporting rubrics suggested in Bock & Mislevy (1988) can present assessment results to these audiences in these modes. We also show how the student-level scores facilitate secondary data analysis.

7.1 State summaries

For media reports, policy debate, and public discussion at the state level, the standard practice is to report achievement in each main subject-matter area on a single scale. The difficulty with a more detailed report is that it may confuse the message that it attempts to convey. Although a multidimensional description of instructional outcomes may be valuable in research, it presents the possibility of

[1] This chapter is based on material prepared for a presentation at a meeting of the American Educational Research Association in New Orleans, April 1988. It is based on preliminary results available at that time. These analyses will be reworked for presentation in the publication version of this report. Figures 7.1 through 7.5 are adapted from Bock & Mislevy (1988).

[2] In addition, item responses and other detailed measures may be reported for .research purposes.

conflicting conclusions about the condition of education in the state. In eighth-grade mathematics, for example, changes in the state mean-scores for the categories of numbers, algebra, geometry, measurement, and statistics will not necessarily be consistent—some may increase while others decline. For purposes of discussion, public officials and the media will want the separate results to be combined into some sort of index that measures the status of overall subject matter. They can then view it with approval or disapproval, as appropriate, and speak univocally about policy initiatives to be taken if difficulties need correction.

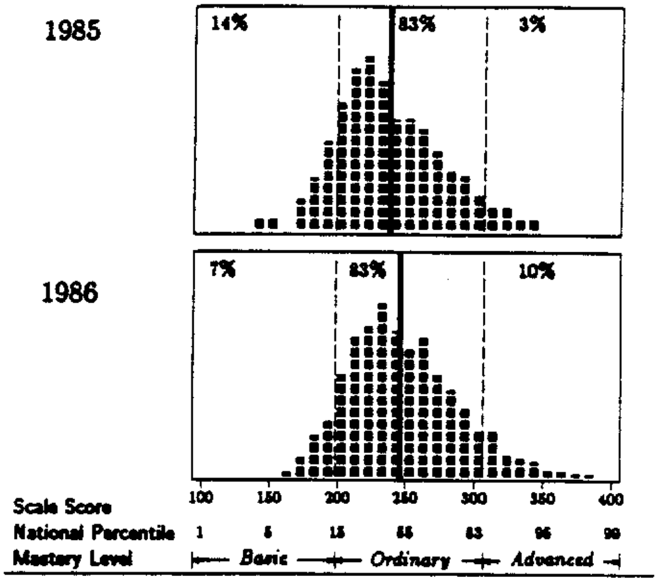
Ideally, these indices should be formed from the topic scores in a way that predicts actual social consequences (for further education, employment, advances in science, *etc.*) of current educational practices. Unfortunately, the validity studies necessary to construct such indices on empirical grounds do not exist at present. Without such studies, we have no alternative but to combine the scales for the separate content areas arbitrarily, often by simple arithmetic averaging of the topic scores to obtain an overall score. This is the approach in reporting the results of the present study. In effect, it implies weighting the areas by the number of items they comprise, since the number of items per area is proportional to the number of topics.

But the reporting of subject-matter attainment as a single scale does not necessarily mean the state report should consist of a *single number*—the average score. The consumers of the state-level report will certainly understand that there is a *distribution* of proficiencies, that some students are performing at high levels of mastery, deserving special opportunities, while others are progressing at a satisfactory but not exceptional rate, and still others are at levels that require remediation. We therefore suggest displaying the state result in the form of a histogram, such as shown in Figure 7.1, which represents the complete distribution of scores. Figure 7.1 shows numbers of students in each 10-point interval of overall student-level mathematics scores. The numbers are either estimated from a sample of schools, as they are in this case, or enumerated from a census of all students at the grade level.

One notices in Figure 7.1 that the distribution is not symmetric, but is skewed toward the higher scores. Whether or not the shape of

STATE SUMMARY

Mathematics scores of 8th Grade Students
in 1985 and 1986



EXPLANATION:
Overall mathematics attainment of 8th grade students in December of 1985 and 1986. Each box (■) represents 1000 students. The heavy line (|) is the median score for each year.

NORC *The University of Chicago*

Figure 7.1. State-level reporting form: distribution of student scores and percent in the mastery groups.

the score distribution can be meaningfully discussed depends upon the manner in which the units of scale are defined. It cannot in general be so discussed when test results are reported in traditional number-right scores; their distributions depend arbitrarily upon the difficulties of the items chosen for the assessment instrument. But Figure 7.1 is based on scale scores computed from the three-parameter logistic models for the items, which, as we have seen, fit the data well. If we accept the simultaneous fit of the models at locations throughout the scale as evidence of comparable units, then the shape of the distribution is meaningful and the appearance of the skew merits explanation.

A possible interpretation is failure of the model to correct fully the effects of guessing on the multiple-choice items. This would put a "floor" on the distribution that would give the appearance of a skew to the right. But we see in Figure 7.1 that the effect seems to involve the whole distribution, and not just the left tail. A better explanation is that the skew is the result of growth of intellectual "capital" which, like the monetary variety, tends to proceed multiplicatively. That is, each step in learning makes the next step easier and quicker, especially in mathematics learning, which is highly cumulative. The outcome of many such independent multiplicative processes is a log-normal form of distribution in the population (see Simon, 1955). Its practical significance is that there is a greater number of students in the state who are especially able in mathematics, and score at very high levels on the test, than would be expected if the distribution were symmetric.

Figure 7.1 is also designed for those persons whose assimilation of information is not helped much by visual displays, or who need to discuss the results in verbal terms. For their benefit, the shape of the distribution is also conveyed by the printed numbers giving percentages of students in three "mastery levels" labeled "Basic", "Intermediate", and "Advanced". (More mastery levels could be defined; NAEP, for example, reports in five levels). In the present instance, these levels were arbitrarily chosen to represent the lower 15 percent, the middle 70, and the upper 15 percent of the estimated state score-distribution.

One would prefer that such levels be related empirically to the educational or vocational potential of students in these score ranges. The policy maker, for example, would like to make such statements as "ninety percent of eighth-grade students who score at or above 280 could be expected to successfully complete a high-school honors program in mathematics," or "ten percent of those below 220 will fail the basic 9-th grade mathematics," *etc.* The mastery levels would then set clear goals for accomplishment in middle school.

In principle, the data required to establish the necessary relationships must exist in district or school records, but they would be accessible for this purpose only if the state's educational information system is capable of following in detail the progress of individual students and summarizing the information in a usable form. Whether many states are in a position to conduct such studies is not clear at present. For those that are not, an alternative might be to equate their assessment scales to those of the National Educational Longitudinal Study (NELS), which has an extensive collection of grade transcripts of students followed from middle through secondary school. A correspondence between criterion points on the NELS test scales could then be established to predict the performance of students in subsequent grades or course work. The NELS data would, of course, apply only very generally to a national sample of students, but the results should be broadly informative in many states.

Still another way to give objective meaning to specific points on the attainment scale is to make use of the property of IRT scoring, mentioned in Chapter 1, that locates the items and the respondents on the same scale. In educational applications, there is some precedent for interpreting points on the scale in terms of the content of items whose 80 percent thresholds are located near a selected point. Students who score at or above that point on the scale then have an 80 percent or better chance of responding correctly to the items near that point. Of course, a given point on the scale is not characterized by any particular item, but by the class of items in the neighborhood of the point. To define a point in a general way, a number of items from the class must be exhibited and their common features inferred. A similar method of interpreting selected mastery levels on assessment scales is used in the reporting of the present NAEP scales in main subject-matter areas.

7.2 Reports for accountability

In states that conduct census assessments, the state educational agencies are in a position to compare the performance of all schools in the state system. The question then arises of how to allow for the impact on student attainment of community factors over which the local education authorities have no control.

Clearly, some adjustment of the school mean-scores for such factors is required if the schools are to be compared fairly. How these adjustments should be made has been widely discussed (for example, in Raudenbush & Bryk, 1989). Two general approaches have been proposed: the first is to express the school means as deviations from the level of performance predicted by measured community factors that are known to be associated with student attainment but are not directly alterable by school officials—typically, linear least-squares regression is used to establish the prediction function; the second method is to perform some sort of cluster analysis on community background characteristics in order to identify groups of "similar" schools—the performance of schools in each cluster is then expressed as deviations from the mean of the cluster.

Inasmuch as the two approaches lead to more or less the same result, we have opted for the somewhat more straightforward regression method; we display the results for the schools in the study in this form in Figure 7.2. Each school in the figure is represented by a code number that would be known to privileged parties. The vertical axis of the graph is the observed mean score for the school; the horizontal axis is the score predicted by the regression equation. In terms of student attainment, schools that are within the upper and lower diagonal lines, which represent two standard deviations vertically above and below the 45 degree identity line, are performing about as expected. Those above the upper line are performing better than expected; those below the lower line, poorer than expected.

Simultaneously, the graph shows on the right-hand scale the absolute performance of the schools, arbitrarily classified as high, average or low. This type of graph overcomes the frequently expressed objections to regression adjustments, that they diminish the school's incentive to improve its position. In the type of graphs shown in Figure 7.2, the poor performing schools, whether or not they are on a par with their peers, are at the bottom of the graph where improvement in absolute terms is obviously needed.

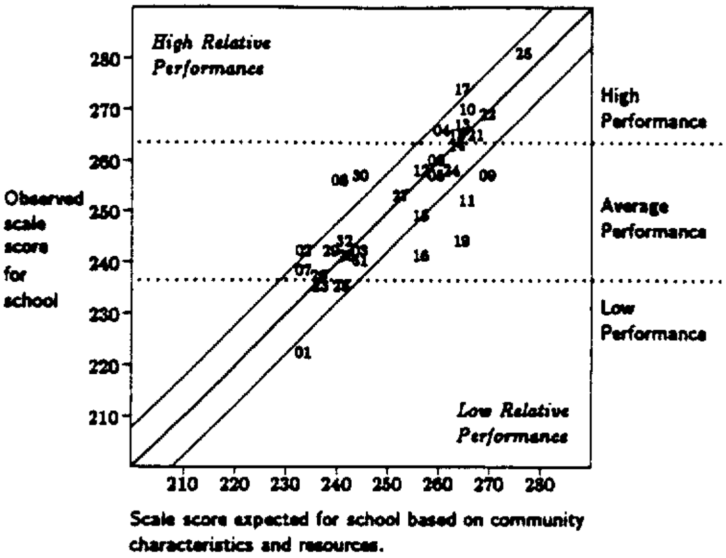
The schools represented in Figure 7.2 are those of the California field trial. The regression model used in the plot is based on the five factors of the allocation sampling design defined in Section 4.1. Each of these variables was dichotomized at the median for the state, and the model included all first-order effects and all two-factor interactions of the resulting variables. All represent conditions over which the schools have little, if any, control.

Another way to make fair comparisons among schools is to examine each school's performance trends from year-to-year. In effect, this is comparing each school with itself rather than competitively comparing one school with another. As we discussed in Chapter 1, these longitudinal interpretations of the assessment depend critically on a high degree of generalizability of the school-level scores and on maintaining a stable reporting scale from one year to another. If these requirements are met, a plot of scores over a period of years should clearly show the general trend for each school, with only minor random deviation about the trend line from such sources as cohort effects, turn-over of school personnel, demographic changes in the community, *etc.* Points in the plot that appear out-of-line would indicate methodological problems—non-standard administration, uncontrolled changes in test forms, compromise of test items, scoring or scaling errors, *etc.*[3] The state agency to which school districts are accountable could maintain "wall charts" displaying these trend lines to help detect problem cases and to monitor the progress of instruction generally. Each chart could be devoted to one subject-matter area, and the lines corresponding to the various schools, organized by district and county. Changes in long-term trends in these charts would alert the agency to potential difficulties in specific schools and districts, and also give evidence of the effects of new programs or school reforms.

[3] The 1986 anomalous NAEP reading results, for example, were found to be due largely to changes in the arrangement of subject-matter content within the test booklets (Beaton, 1988).

STATE SUMMARY

School Performance Chart
8th Grade Mathematics



EXPLANATION:

The location of schools on the performance chart is indicated by their identification codes. *Absolute* performance levels are given by the positions of the schools on the scale on the left. Performance *relative* to other schools with the same community characteristics and resources is indicated by the vertical distance of the school code from the heavy diagonal line. Schools located above the upper light diagonal line are performing better than expected. Those below the lower light diagonal line are performing less well than expected.

NORC *The University of Chicago*

Figure 7.2. State-level reporting for: observed versus expected performance of the schools.

7.3 School reports

We suggest two types of school reports, one designed to monitor the distribution of attainment among the students, and the other to evaluate progress toward curricular objectives. The reports combine features of graphical and numerical presentation. They are designed for distribution to school principals, with copies to the district superintendent.

An example of the first type of report is shown in Figure 7.3. The data are those of a typical school in the California field trial (the name of the school has been changed). The distribution of student-level scores for each of the main content areas and each of the process proficiencies is shown in intervals of 16.66 scale points, and each dot represents about 3 students in this school. (In larger schools, each dot would represent a larger number of students.) The heavy vertical bar represents the school's overall mathematics score, *i.e.*, the score that determines the location of the school in the state distribution shown in Figure 7.1. The corresponding numerical value is shown below the graph, along with the equivalent percentile point in the score distribution of Figure 7.1. For each content area and proficiency, the score distribution is also characterized by the school mean, the corresponding state percentile, and the percent of students at each of the three mastery levels.

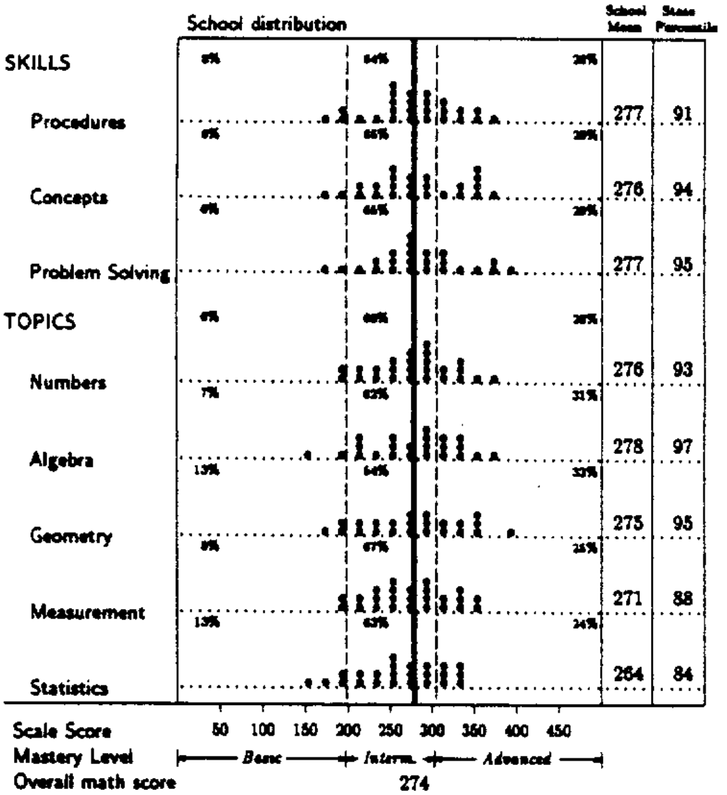
A particularly effective use of the information in this part of the school report would be an annually updated graph showing attainment in each subject-matter area. Departments within schools could have similar graphs detailed by content topic and proficiency in their particular subject matter. Prominently displayed, the progress of the school represented in these various graphs would be a source of interest and motivation to students and staff alike.

If the students have a relevant classroom assignment, a report similar to that in Figure 7.3 can be provided to the classroom teacher. In these reports, each dot would represent the location of a particular student in the classroom distribution. If the teacher is also provided a roster of scores for all students in the classroom, he or she can identify locations of students on the graph by name for purposes of student guidance or for conferences with parents. In this role, the classroom

Survey Test of Grade 8 Mathematics

SCHOOL REPORT

School: Sanderson
Date of Testing: 11-11-86
Number of Students Tested: 72



	School Mean	State Percentile
SKILLS		
Procedures	277	91
Concepts	276	94
Problem Solving	277	95
TOPICS		
Numbers	276	93
Algebra	278	97
Geometry	275	95
Measurement	271	88
Statistics	264	84

Overall math score 274
State Percentile 94

EXPLANATION: Each • represents about three students. The heavy black vertical line marks the overall average score of the school in mathematics.

NORC *The University of Chicago*

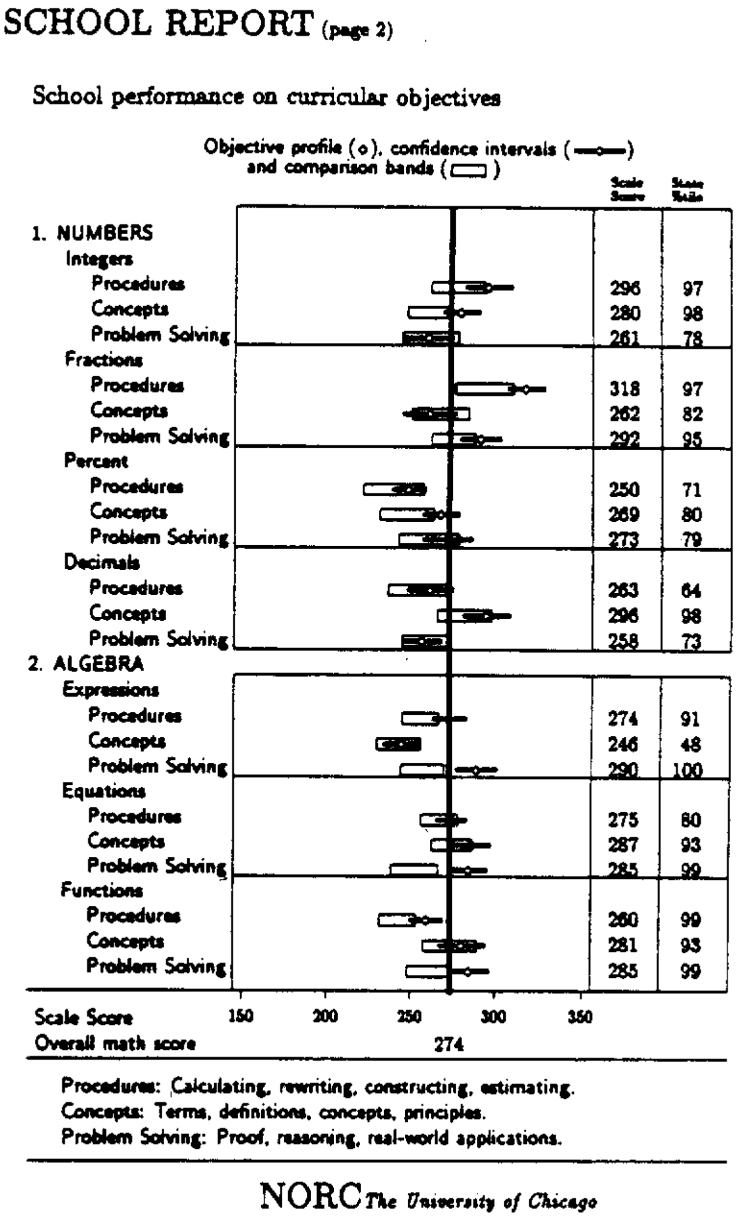
Figure 7.3. School-level reporting form: distribution of student scores in main skill and content areas.

report supplements the individual-student reports described in Section 7.4.

The second type of school report, shown in Figure 7.4, is intended primarily for the evaluation of instruction. Displaying a school-level score for each cell in the content-by-process classification of the assessment exercises, the report depicts progress in curricular objectives at the topic and skill level. School-level scores based on a 45-minute test cannot measure the objectives in greater detail than with this acceptable degree of generalizability. Although more detail would be available for very large schools if more than one Duplex-structured instrument were administered simultaneously to different students, many schools would not have enough students at a grade level to make this feasible. Moreover, the student-level scores for students administered different instruments would not be as rigorously comparable as scores based on the same instrument, as is the case in the present study. For these reasons, we limited the mathematics assessment to the forty-five objectives shown in Figure 7.4.

In Figure 7.4, the school's overall performance at grade level is again represented by the heavy vertical line at the same scale point as in Figure 7.3. As explained in Chapter 6, the units of these school-level scores are adjusted so that they can be expressed on the same scale as the student-level scores in Figure 7.3. Because they represent the average performance at grade level, the variability of the values shown in this report is less than that of the distribution of student scores in Figure 7.3. This is the reason that the range of the scale in Figure 7.4 is smaller than that of 7.3.

The scores for the school are shown as a profile about the line for the overall score. The value of the score for each objective is represented by the diamond, and it is also given at the right as a numerical value and as a state percentile in the distribution of *school-level* scores (not a percentile of the *student*-score distribution). Each of the diamonds is bordered by a bar indicating a one-standard-deviation confidence interval on the true score. This interval has a probability of about two-thirds of including the true score for the school. Its length is a function of the number of items in the matrix sample for the corresponding objective, the average discriminating power of those items, the score level, and the number of students at the grade level.



SCHOOL REPORT (page 2)

School performance on curricular objectives

Objective profile (○), confidence intervals (—○—)
and comparison bands (▭)

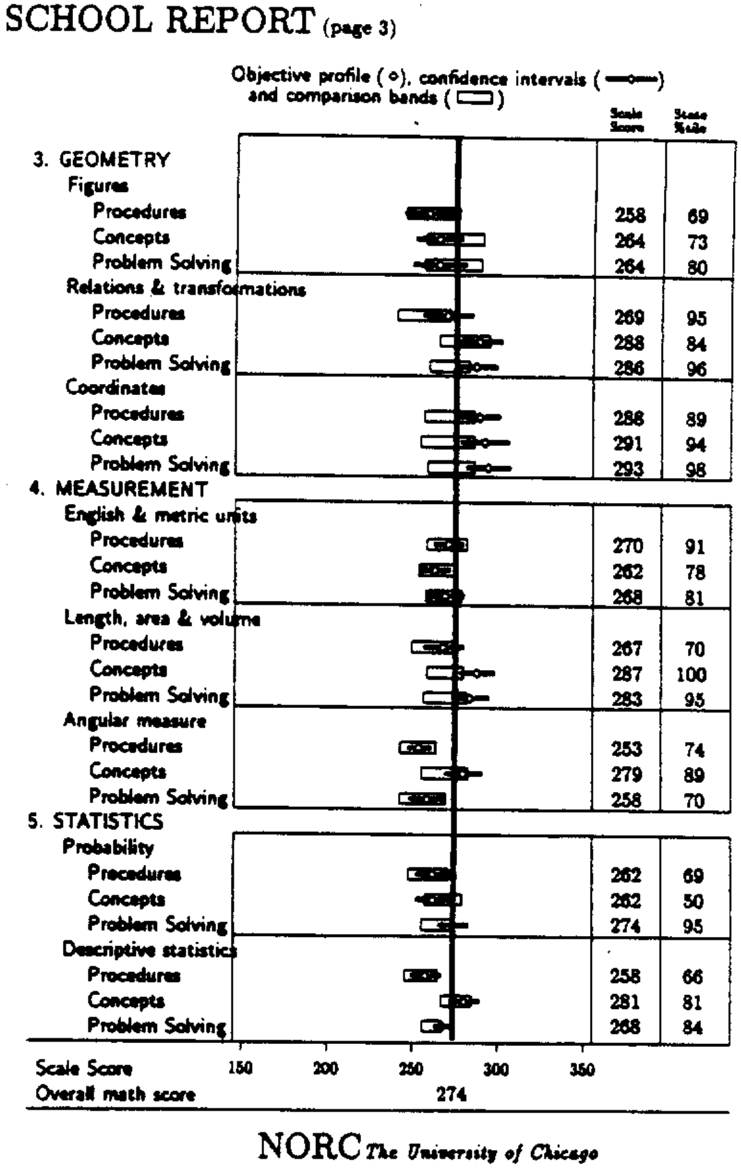
	Scale Score	State %ile
1. NUMBERS		
Integers		
Procedures	296	97
Concepts	280	98
Problem Solving	261	78
Fractions		
Procedures	318	97
Concepts	262	82
Problem Solving	292	95
Percent		
Procedures	250	71
Concepts	269	80
Problem Solving	273	79
Decimals		
Procedures	263	64
Concepts	296	98
Problem Solving	258	73
2. ALGEBRA		
Expressions		
Procedures	274	91
Concepts	246	48
Problem Solving	290	100
Equations		
Procedures	275	80
Concepts	287	93
Problem Solving	285	99
Functions		
Procedures	260	99
Concepts	281	93
Problem Solving	285	99

Overall math score 274

Procedures: Calculating, rewriting, constructing, estimating.
Concepts: Terms, definitions, concepts, principles.
Problem Solving: Proof, reasoning, real-world applications.

NORC *The University of Chicago*

Figure 7.4. School-level reporting form: attainment of curricular objectives.



SCHOOL REPORT (page 3)

Objective profile (○), confidence intervals (—○—) and comparison bands (▭)

	Scale Score	State %ile
3. GEOMETRY		
Figures		
Procedures	258	69
Concepts	264	73
Problem Solving	264	80
Relations & transformations		
Procedures	269	95
Concepts	288	84
Problem Solving	286	96
Coordinates		
Procedures	288	89
Concepts	291	94
Problem Solving	293	98
4. MEASUREMENT		
English & metric units		
Procedures	270	91
Concepts	262	78
Problem Solving	268	81
Length, area & volume		
Procedures	267	70
Concepts	287	100
Problem Solving	283	95
Angular measure		
Procedures	253	74
Concepts	279	89
Problem Solving	258	70
5. STATISTICS		
Probability		
Procedures	262	69
Concepts	262	50
Problem Solving	274	95
Descriptive statistics		
Procedures	258	66
Concepts	281	81
Problem Solving	268	84

NORC *The University of Chicago*

Figure 7.4 (continued). School-level reporting form: attainment of curricular objectives.

In this application, only the two latter factors vary between schools, so the length of the intervals depends entirely on the level of the score and school size. Scores near the center of the distribution of schools tend to have smaller confidence intervals than those at the extremes.

The report in Figure 7.4 has another interpretive feature, borrowed here from the school reports of the California Assessment Program. The score for each curricular objective is accompanied by a so-called "comparison score band" represented by an open rectangle. The comparison score band is the predicted score for the school (computed for each objective from a regression equation based on the same community background factors used to compute the expected overall mathematics score in Figure 7.2) plus or minus one standard deviation of the sum of the residual variation in prediction and the measurement error variation of the school score. The probability that the band includes the school score when the background factors account for it entirely is therefore about two-thirds. Thus, when the comparison band includes the diamond, the school is performing about as expected in this objective, but if the diamond is to the left of the band, there is some indication that instruction for the objective is below the community norm. Conversely, if the diamond is to the right of the band, there is evidence that the outcome of successful instruction is above expectation.

An interesting question is whether the comparison score band should be recomputed every year. If it is, the school will have the impression of not making progress in conditions where instruction is improving in all schools in the state. The same misleading impression would be conveyed by the state percentiles. A case could therefore be made for keeping the prediction models fixed for a number of years so that schools could compare themselves relative to the base-year expectations. (The same reasoning could be applied to the state summary in Figure 7.2.) Or, alternatively, equal emphasis should be given to trends in the absolute scores for the objectives by providing, the kind of longitudinal graphs that we have recommended for the school-average of the student-level scores.

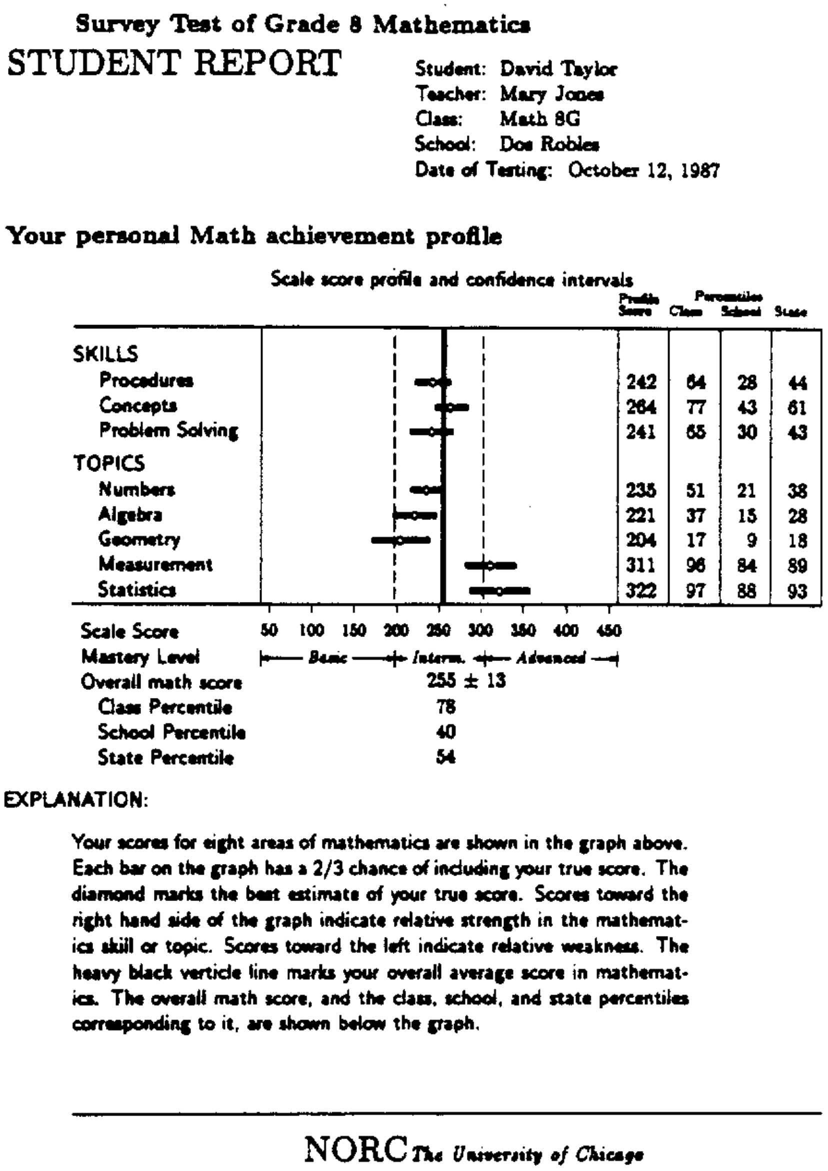
The particular school shown in Figure 7.4 is interesting in that the students are performing better than expected for problem solving in algebra and geometry, which are objectives not always well-represented in instruction at the eighth-grade level. This may mean that the school has advanced classes in these topics taught by someone who emphasizes work on problem solving.

7.4 Student reports

Our suggested form for the student report, shown in Figure 7.5, depicts the performance of a student in the California field trial (the name of the student, the teacher and the school have been changed). Copies of this report could be laser printed for the student and parents, for a teacher or counselor, and for the student's folder. Or it could be part of a more elaborate report showing each student's status at the end of the school year.

As we pointed out in Chapter 5, the eight distinct dimensions of mathematics achievement represented in Figure 7.5 are about the most that can be reliably reported from a 45-item test, even with two-stage testing. They contain enough information to assist student guidance and placement, and parent counseling, but they should be used for these purposes only in combination with classroom test results, productions in the student's folder, and teachers' or counselors' observations.

The heavy vertical line in Figure 7.5 indicates the overall mathematics score of the student in question. Scores for the three process proficiencies, labeled "skills" in the figure, and for the five main content areas, called "topics", are presented on an absolute scale with mean 250 and standard deviation 50 in the state distribution of respective student-level scores. Displayed as a profile about the vertical line for overall performance, each score is marked by diamonds and also presented numerically at the right. The plus or minus one-standard-error-band on the score is the heavy horizontal bar containing the diamond. The bar has a probability of about two-thirds of including the student's true score (*i.e.*, the score that the student would obtain if tested on indefinitely many items from the respective skill and content domains). We suggest using the bars to interpret the student's relative strengths or weaknesses in particular areas relative to his or her overall mathematics performance. If the band includes the verti-



Survey Test of Grade 8 Mathematics

STUDENT REPORT

Student: David Taylor
Teacher: Mary Jones
Class: Math 8G
School: Dos Robles
Date of Testing: October 12, 1987

Your personal Math achievement profile

Scale score profile and confidence intervals

	Profile Score	Percentiles Class	School	State
SKILLS				
Procedures	242	64	28	44
Concepts	264	77	43	61
Problem Solving	241	65	30	43
TOPICS				
Numbers	235	51	21	38
Algebra	221	37	15	28
Geometry	204	17	9	18
Measurement	311	96	84	89
Statistics	322	97	88	93

Scale Score 50 100 150 200 250 300 350 400 450
Mastery Level ⊢— Basic —+— Interm. —+— Advanced —⊣

Overall math score 255 ± 13
Class Percentile 78
School Percentile 40
State Percentile 54

EXPLANATION:

Your scores for eight areas of mathematics are shown in the graph above. Each bar on the graph has a 2/3 chance of including your true score. The diamond marks the best estimate of your true score. Scores toward the right hand side of the graph indicate relative strength in the mathematics skill or topic. Scores toward the left indicate relative weakness. The heavy black verticle line marks your overall average score in mathematics. The overall math score, and the class, school, and state percentiles corresponding to it, are shown below the graph.

NORC *The University of Chicago*

Figure 7.5. Student-level reporting form: achievement profile in skill and content areas.

cal line, no inference of either strength or weakness is made. If it is entirely to the left of the line, the inference is that the student should be able to improve in that area by extra study. If the bar is to the right of the line, it may indicate that the student has particular accomplishments in certain parts of the subject matter. The student in Figure 7.5, for example, may have special interests in empirical science that give him an advantage in Measurement and Statistics relative to the more formal content in Numbers, Algebra and Geometry. He is unusually poor in Geometry, even relative to his classmates.

If the assessment covers several subject-matters and the students are assigned to different classrooms for each, it would not be practical to report class percentiles as we have done here. To do so would require information in computerized form on all the classroom assignments of every student; few schools at present maintain data bases with this capability, although it might exist at some time in the future.

School and state percentiles would always be available, however, and these would allow the scores to be interpreted both relative to a community standard and to a state standard. If the procedures were implemented for equating the state assessment to a nationally-normed assessment, such as NAEP, national percentiles could also be reported. In the absence of empirically established predictive criteria to which the student level could be referred, the normative percentiles are only aids to interpretation that are available when the assessment is first introduced. If the assessment scale is kept consistent over a period of years, teachers and other persons who regularly see these reports will begin to attach a more absolute meaning to the scores. Knowing where particular students were located on the scale, they will begin to generalize about the implications of the scale values for the ability and potential of students moving through their school. Normative scales that have been in use for a long period, such as those of the Scholastic Aptitude Test (SAT), have attained this status.

As we mentioned in connection with the state percentiles for the school scores in Figure 7.3, there is a question as to how frequently the normative base should be updated. We argued in the case of schools, where change is seen only over a period of years, that the percentile norm should refer to a base year fixed for some period, perhaps five or ten years. But in the case of the individual-student scores, the institutional perspective is not as relevant, and updating to the current assessment seems reasonable. The percentiles would then show each student's rank in his or her cohort, and to some extent this ranking would retain its relevance as the cohort moves onward in education or careers.

It is important to understand, however, that the *mastery levels* indicated by the vertical dashed lines in Figure 7.5 are content-referenced, not norm-referenced, and would not change from year-to-year. Neither are they predictive criterion levels based on prospective studies of students' careers. As we have discussed in Chapters 1 and 3, and in Section 1 of this Chapter, they refer rather to the probability of success in the typical content of the assessment tasks and exercises whose 80 percent thresholds fall in the vicinity of these lines. As a further aid to interpretation, some description of the relevant item content should accompany the student-level reports.

We are indebted to Richard Hill for pointing out that a further type of student report may be desirable. As an aid to counseling students and parents, a brief computerized record of the response alternatives marked by each student should be returned to the school counselor. The booklet number would be indicated, and each item response would also be scored "right" or "wrong". The counselor, who would be provided with a complete set of the test booklets, could then discuss, with students or parents who might be interested, the particular items that gave the student difficulty. In this way, the parents could be relieved of any feeling that their child was evaluated on a test of unknown nature and content.

7.5 Dimensionality of the student-level scores

The eight scores of the student profiles in this study were chosen for their relevance to the mathematics curriculum and to current cognitive theories of mathematics learning. The choice does not necessarily imply, however, that the mathematics attainments of eighth-grade students actually vary in this many dimensions. Nearly all cognitive proficiencies are fairly highly correlated across a general population, and the partial information gained as additional dimensions of performance are added decreases rapidly. The conventional approach to investigating the dimensionality of individual differences in psychological-test scores is some form of factor analysis that includes a statistical criterion of the number of significantly resolvable dimensions. Psychologists who have attempted to define human intelligence empirically—Thurstone, Guilford, Cattell, and others—have used this method with some success. To apply it in the present study is more difficult, however, because the Duplex instrument consists of multiple, similarly structured forms administered to different respondents. In place of a conventional maximum likelihood factor analysis, for example, some form of multiple-group analysis, such as that provided by the LISREL program of Jöreskog & Sörbom (1989), would be required.

But whether such a factor analysis of individual differences is germane to assessment measures is not entirely clear. The variation that is observed in instructionally relevant variables depends not only upon the fundamental dimensions of human capability, but also on differences in what is being taught or how well it is taught in different parts of the population. These latter influences are subject to change according to policy or consensus; no one analysis at any point in time would definitively establish the dimensionality of instructionally sensitive measures. Such an analysis would commit the "existential fallacy"—the inappropriate identification of what *is* with what *can* or *ought* to be (see Stodolsky, 1988).

It is also well-known that the factor analysis of individual differences is not sensitive to the group-level effects that are of interest to someone investigating curricular or instructional effects. To investigate the dimensionality of these effects, one wants a discriminant analysis rather than a factor analysis. Applications of discriminant analysis to educational data are described in Bock (1966, 1975), Finn (1974), and other texts on multivariate statistics. The MULTIVARIANCE program of Finn & Bock (1989) includes especially complete facilities for discriminant analysis and the associated statistical tests of between-group dimensionality.

As an illustration of this method of investigating the dimensionality of instructional effects, we have performed discriminant analyses of between-classroom and between-teacher variation in some of the larger schools in the California field trials. These analyses estimate linear functions of the scores that maximally discriminate between 8th-grade mathematics classes in these schools. Results of two of the more interesting analyses (for school No. 18 and school No. 30 in Figure 7.2), in which separate analyses are carried out for the three proficiency variables and the five content variables, are discussed here. The summary statistics for these analyses are shown in Tables 7.1 and 7.2. The mean scores are shown for classrooms identified by teacher and Honors or Advanced Placement status.

In school No. 18, the discriminant analysis of the ***process proficiency*** measures revealed two statistically significant dimensions among the means of the fourteen classrooms. The corresponding standardized discriminant functions were:

$$\begin{aligned} V_1 &= 0.507Y_1 + 0.644Y_2 - 0.094Y_3 \\ V_2 &= -1.263Y_1 + 0.350Y_2 + 1.189Y_3 \end{aligned}$$

where Y_1, Y_2, and Y_3 are the Procedural Skills, Conceptual Understanding, and Problem Solving scores, respectively. The first function can be characterized as a Procedures + Concepts variable, and the second as a contrast between Problem Solving and Procedures (Problem Solving − Procedures). Labeled in this way, the mean discriminant values (group centroids) of the classrooms are plotted as shown in Figure 7.6.

An interesting result in this plot is that, except for the honors classrooms and the apparent remedial classroom taught by teacher A, there is clear clustering of classrooms taught by the same teacher. This means that the teachers had a consistent "style" relative to their emphasis on procedures and concepts as opposed to problem solving. The classes of teacher C, for example, were at about the same level as their counterparts on the Procedures and Concepts dimension, but were considerably lower in the Problem Solving minus Procedures dimension (*i.e.*, they are relatively poorer in problem solving.) Indeed, the honors class of teacher C was at the same level on the latter dimension as the regular classes—a situation that should be of some concern for mathematics instruction in this school, given that reasoning and problem solving should have a special place in the honors curriculum. The classroom means of the other teachers were much more homogeneous, but their tendency to cluster by teacher is still

TABLE 7.1

Summary Statistics: Classrooms in School No. 18

Class Means				Proficiency Measures		
Class	Teacher	Honors	N	Procedures	Concepts	Problem Solving
1	A		22	199.7	205.1	214.1
2	A		22	266.4	257.9	249.1
3	A		26	247.3	250.3	244.8
4	B	*	38	317.8	318.3	301.0
5	B		23	266.4	261.8	257.0
6	B		24	262.0	257.3	255.0
7	C		22	258.3	272.1	276.0
8	C		20	250.8	253.6	272.8
9	C	*	28	312.2	319.4	310.6
10	D		26	260.5	259.6	258.4
11	D		23	270.5	272.4	268.6
12	E		23	254.5	244.3	244.7
13	E		19	246.4	234.8	246.3
14	E		24	255.7	258.2	254.7
Common within-class standard deviations: df=36				37.9	37.7	40.7
Correlations						
				1.000		
				.710	1.000	
				.684	.661	1.000

TABLE 7.2
Summary Statistics: Classrooms in School No. 30

Class Means				Proficiency Measures				
Class	Teacher	Honors	N	Numbers	Algebra	Geometry	Measurement	Statistics
1	A		16	207.0	211.9	224.6	211.5	216.8
2	A		14	187.2	209.1	217.9	209.8	191.2
3	A		15	231.1	234.4	246.9	230.0	232.6
4	B		24	227.5	234.7	242.5	232.0	224.7
6	C		9	231.7	228.7	234.0	224.6	222.1
7	C		18	217.2	226.9	226.8	224.9	210.2
8	D		25	193.0	199.6	196.6	198.6	196.2
9	D	*	32	275.4	276.3	266.8	259.2	259.1
10	D	*	32	285.5	287.7	280.3	289.3	269.3
11	E		15	220.5	217.5	250.5	230.4	223.3
12	E		19	204.5	223.6	227.0	219.8	228.5
13	F	*	19	290.9	313.3	265.3	279.9	264.4
14	G		24	181.3	193.3	184.5	186.8	187.3
15	H		27	241.4	233.8	231.3	225.2	210.1
16	H		21	211.2	216.9	230.9	209.9	214.0
17	H	*	17	275.5	279.2	265.8	260.8	264.2
18	I		24	220.8	218.5	216.3	223.3	222.5
21	I		23	248.4	249.4	249.1	237.8	233.3
22	J		25	237.5	240.5	238.2	237.4	233.8
23	J		25	209.3	215.8	205.8	216.2	207.4
24	J		13	215.6	217.5	212.3	206.5	211.5
Common within-class standard deviations: df=416				37.4	38.5	40.4	38.9	37.2
Correlations								
				1.000				
				.533	1.000			
				.478	.469	1.000		
				.610	.580	.488	1.000	
				.540	.493	.479	.585	1.000

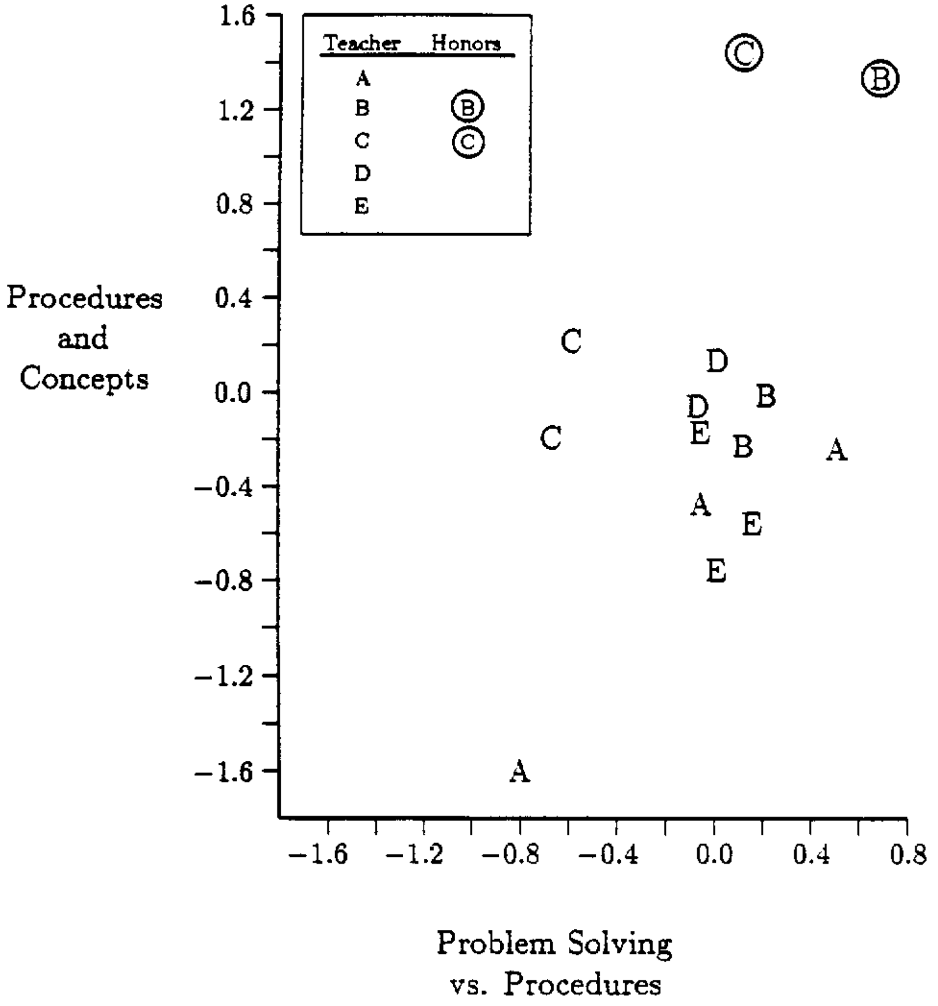


Figure 7.6. Discriminant analyses of teachers and classrooms in School No. 18: Student-level proficiency measures.

apparent.

A second discriminant analysis, for the very large school No. 30, showed two significant dimensions in the classroom means for the five *content* variables. The standardized discriminant functions defining these dimensions were:

$$\begin{array}{rcrcrcrcrcr} V_1 & = & 0.563Y_1 & + & 0.461Y_2 & + & 0.101Y_3 & - & 0.003Y_4 & + & 0.073Y_5 \\ V_2 & = & -0.560Y_1 & - & 0.641Y_2 & + & 0.931Y_3 & + & 0.321Y_4 & + & 0.369Y_5 \end{array}$$

where Y_1 through Y_5 are the Numbers, Algebra, Geometry, Measurement, and Statistics content scores, respectively. Apart from the very small negative coefficient for measurement, the first function is a sort of overall mathematics variable, with emphasis mostly on Numbers, Algebra, and Geometry. The second is a contrast between content that depends primarily on symbolic presentation—Numbers and Algebra— and content in which graphical presentation plays a larger role—namely, geometry, measurement, and statistics.

The plot of the mean discriminant values for the classrooms, identified by teacher and honors status, is shown in Figure 7.7. Again there was a tendency for the classroom means to cluster by teacher. Teacher E was especially effective in the graphical content; while his or her classes were below average in the symbolic content, they were in the range of the honors classes in the graphical content. The honors class of teacher F was very high on the symbolic content, but only middling on the graphical content. Teacher G must have had a remedial class, or was not a very effective instructor; but because he or she was teaching only one class, the two possibilities cannot be distinguished.

These analyses illustrate the interesting results available to secondary analysis when the assessment design provides, in addition to group-level scores, good-quality measures at the student level. The data from state assessment programs, especially the census assessments at multiple grade levels, then contain information in such abundance that a great range of educational research hypotheses can be investigated. Regrettably, there has been relatively little effort to exploit this potential of state assessment data for secondary studies. The problem is partly that the data have not reached the right investigators, persons familiar with the analysis of multilevel data (see

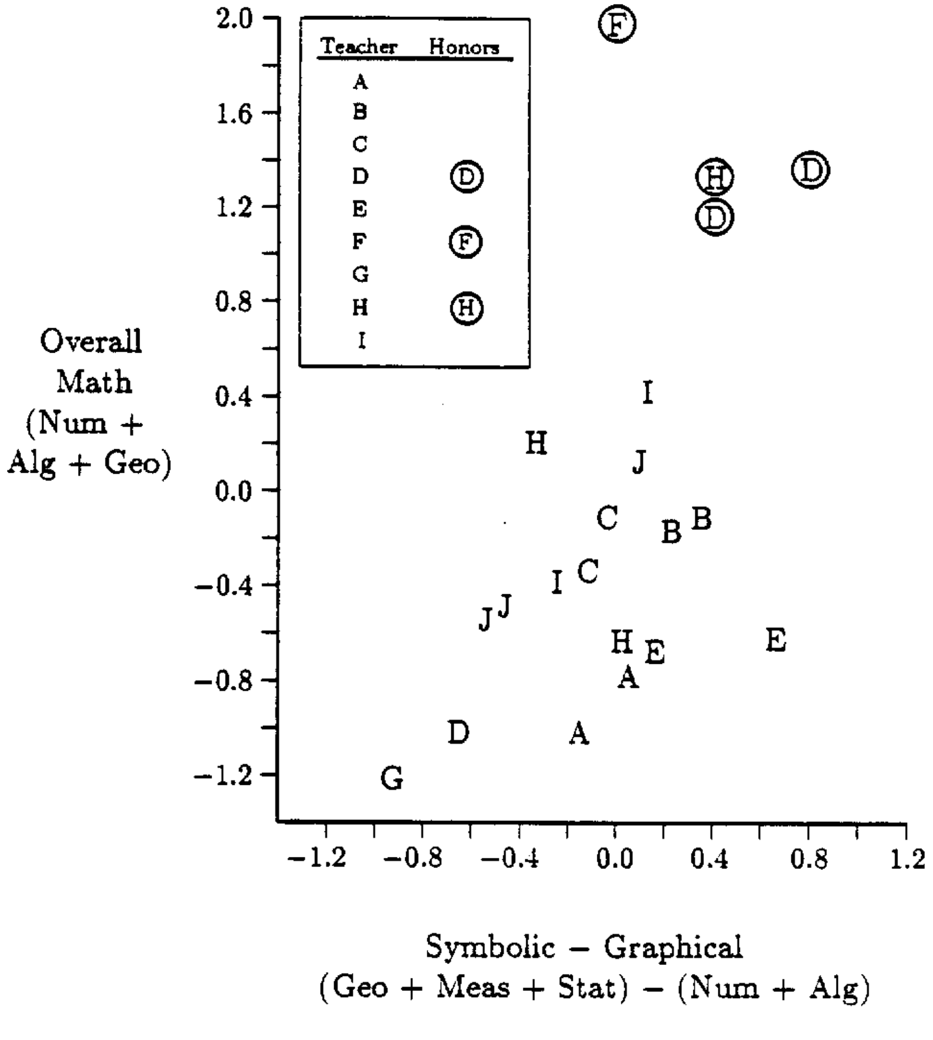


Figure 7.7. Discriminant analyses of teachers and classrooms in School No. 30: Student-level content measures.

Bock, 1989), but partly also because the form or quality of the data has not been suitable for such studies. A particular strength of the Duplex Design is its ability to provide, simultaneously with the operational information required for policy and management purposes, a form of data suitable and convenient for secondary research.

7.6 Summary

The impact of the assessment on the educational policy of the state, on the shaping of the curriculum and management of instruction, and ultimately on the performance of students, depends upon the communicating power of its reports. A guiding principle of the reporting should be that, insofar as possible, the data should speak for themselves in easily understood displays. The consumers of the reports should have the opportunity to examine the data summaries and judge the condition of student learning from their own perspective. The assessment agency would do well to maintain a degree of objectivity and avoid involvement in potentially controversial interpretations of the results. Its job should be to provide dependable information in the clearest and most usable forms, and not to prejudge the educational, policy or scientific implications of the findings.

In some situations, however, the reporting cannot avoid becoming involved in interpretational questions. The prime example is the adjustment of scores to account for the effects on student attainment differences in community characteristics over which the local education authorities have no control. Because a commitment to some theory of community influence on student attainment is implied in these adjustments, their use carries the assessment beyond the realm of purely objective reporting. But there is no obvious way of avoiding such adjustments, or similar treatment of the data, if results from districts and schools in widely differing community settings are to be compared fairly. We have tried to preserve as much objectivity as possible in this connection by presenting the data (in Figure 7.2) in a display from which both the adjusted and the unadjusted results can be seen simultaneously. But ultimately the credibility of the comparisons will depend on progress of research in clarifying the relationships between community background factors and the performance of students in the schools.

For many purposes, graphical presentations of results convey more information in a more accessible form than the myriad tables one associates with official educational statistics. We have illustrated here some of the display formats that have been proposed for reporting assessment results to various audiences. We attempt to convey not only the average levels of performance, but also some indication of the range and distribution of outcomes in the relevant population or group. Where possible, the stability of the results is indicated by confidence intervals for the mean values reported. Recognizing that for purposes of verbal discussion, numerical values are also necessary, we have also included the various quantities in the margins of the display, both as scores on the defined assessment scales and as percentiles of relevant populations.

While accepting that for policy and media uses the assessment results have to be abstracted in simple one-dimensional social indicators, we realize that anyone directly involved in the day-to-day work of education knows that learning is inherently multidimensional. Changes in numerous areas must be monitored simultaneously; otherwise, gains in some areas may be at the expense of losses in others. For this reason, we have emphasized the use of reporting profiles—some highly detailed for the evaluation of curricular outcomes—others less detailed for the guidance of individual students. The profiles convey the patterns of relative strengths and weaknesses that can be diagnostic of unevenness in instruction or unsteady application of students.

In the remaining important function of assessment—providing data for research studies of educational phenomena—the method of displaying the results is less important than their quality and their availability in a well-documented, machine-readable form. The mandates of state assessment programs do not ordinarily include making data available for secondary analysis, but weighed against their potential contribution to educational research, the cost of their distribution for investigative use is small. We have therefore emphasized in the implementation of the Duplex Design that assessment procedures not only deliver dependable data for policy, management, and guidance purposes, but also produce dependable data for research. We illustrated this feature of the design by applying student-level scale-scores from the California field trial in discriminant analyses of classroom and teacher effects within large schools. We found clear evidence for teaching "styles" in more than one dimension of both the content and process measures. These analyses demonstrate in a small way the potentially interesting and informative studies that could be carried out with the data base of a well-designed assessment program.

Chapter 8

Summary and conclusions

Many features of present educational assessment programs originated in the early development of the National Assessment of Educational Progress. Perhaps the most important was the emphasis on creating a continuous record of average student attainment by which long-term changes in the effectiveness of American education could be judged. Placing the emphasis on *change* made it clear whether progress was favorable (positive change) or unfavorable (negative change) even when the absolute levels of the scores could not be interpreted. It also encouraged the use of graphs to display a time-series of assessment outcomes over a period of years, thus showing vividly the direction in which American education was moving.

Another feature of the national program was the desire for a comprehensive assessment, not confined just to the "three-R's", but encompassing a much wider domain of subject matter and extending even to traditional creative skills not strongly tied to subject matter. Early assessments included such topics as vocations, and basic skills in art and music. Only recently, with the new emphasis on performance assessment, discussed below, have these areas again been considered for evaluation.

There was also an expectation that the assessment would serve many purposes, from describing the state of our educational effort to a broad public audience, to guiding national educational policy, evaluating federal initiatives to improve instruction, and following trends in attainment in major regional areas and sociocultural groups. Although the national assessment only slowly began to make an impression on the education community, the growing national awareness of our need for an informed and capable citizenry in a time of international economic competition worked in its favor. Stimulated also by the comparisons of student performance in other nations carried out by the International Educational Achievement Association, concern for the state of education revealed by the U.S. assessment took on a new urgency. A movement among the chief state school officers to expand the scope and functions of the national assessment led to Congressional support for extending the data collection to allow comparisons between states.

Finally, the framers of the national assessment program also intended it to provide innovative procedures and exercises to state testing programs for use in locally developed instruments, as well as to accumulate a rich store of data for independent investigators to use in the study of educational and societal problems. In the first of these roles, it was quite successful. Not only were items from the national assessment widely used in other testing programs, but the new approach it introduced in data collection set the trend for many of the state assessment programs.

8.1 Matrix-sampling designs

In the earliest planning for the National Assessment of Educational Progress, it was recognized that conventional achievement tests would not serve all these purposes. Because they are designed to provide accurate scores for individual students, such tests are too long and time-consuming to employ in a national testing effort, even one with a low sampling rate. Nor do they include enough items to cover many facets of the curriculum or to allow some fraction of the items to be retired after each assessment for public disclosure or distribution to the states. But even more important, such tests typically exist in only a few forms and do not have a high degree of generalizability to the content and skill domains they purport to measure. The interaction of their limited item representation with the changing samples of schools and students from one assessment year to another produces too much temporal variation in the measures of attainment among the subgroups of interest in the population.

For all of these reasons, the national assessment adopted matrix-sampling designs as their basic method for collecting information on student performance. In these designs, any given student sampled from some school and classroom responds to only a relatively small number of items sampled from the content and skill domains. These item samples appear in the multiple test forms and booklets that make up the assessment instrument. The booklets are assigned to students in such a way that all students have the same probability of receiving any given form. This allows the item responses to all forms to be aggregated when estimating average scores for various subgroups of students. If the large total set of items contains representatives of various curricular objectives, each objective can be accurately assessed at the group level.

Matrix-sampling designs have been the key to efficient and economical measurement of long-term trends in student attainment by national and state assessment programs. They have a number of advantages that encourage their use. In particular, the large number of items that comprise the instrument discourages any attempt to teach the answers to particular items. The sheer number of items also protects the assessment program from the accidental exposure or compromise of some of the items. The affected items can be retired and easily replaced without altering comparability of the measures. Indeed, when item response theoretic (IRT) methods are used to score the assessment results, a certain proportion of the items can be renewed at frequent intervals, thus keeping the content up-to-date and guarding against excessive item exposure. The IRT methods make it possible to perform this type of item updating using only the data obtained in the regular, operational assessment. Only more major changes in the composition or format of the assessment instruments require so-called "bridge studies", in which old and new forms are randomly assigned to schools, or preferably, students within schools, in order to equate the instruments and keep the time series of assessment results comparable between the old and the new instrument.

Another advantage of matrix-sampling designs is their suitability for special studies and evaluations of instructional programs. Because of the many forms, the assessment instrument can rather freely be given to the same students before and after instructional treatments.

Typically, this would mean autumn and spring testing of classrooms taught with different methods or materials. There is only a small probability that a given student will receive the same form in both testings, and even for those students who do, the adjustment required to account for the pre-exposure effect is easy to estimate (provided the assessment forms are distributed randomly to the students on each occasion.)

While matrix-sampling instruments enjoy these many advantages in large-scale assessment applications, they suffer from one major disadvantage: they do not provide student-level scores. We have pointed out in Chapter 1 the many undesirable consequences that follow from this limitation. Because they are capable of reporting only at group levels, such as the program, school, or school system, these designs can not provide the information about individual students that is the main concern of classroom teachers, the parents, or the students themselves. Moreover, in failing to provide case-by-case measures of attainment, they do not provide data in a form that most researchers need in order to study the relationship of student characteristics to attainment or to perform multilevel analyses of student and school effects.

But perhaps most important, in the policy and public information uses of assessment results, the group-level scores alone do not inform discussion of educational problems in the language most readily understood by general audiences—namely, in terms of numbers or percentages of students whose achievement meets, or does not meet, the standards required in a society that needs increasing numbers of well-educated young people.

These shortcomings of matrix-sampling designs are at present limiting the contribution that assessment makes to educational planning and improvement. Their correction lies not in abandoning matrix sampling, but in developing it into a more powerful type of design that provides group-level and student-level scores simultaneously and efficiently. Such a design now exists.

8.2 The Duplex Design

It was to overcome the chief disadvantage of matrix-sampling designs, while preserving their advantages, that Bock and Mislevy (1988) proposed the Duplex Design. By the special structuring of the item content within the test booklets, the Duplex Design yields informative and dependable scores for individual students in main content and skill areas without interfering with scoring across the forms to measure progress in detailed curricular objectives at the school and higher levels.

In the present studies, supported by the Office of Educational Research and Improvement (OERI), U. S. Department of Education, and coordinated by the Center for Research in Evaluation, Standards, and Student Testing (CRESST), UCLA, we undertook to test the feasibility of this new type of assessment design. With the cooperation and help of the state assessment programs in Illinois and California, and the good offices of the respective chief state school officers, Ted Sanders and Bill Honig, we were able to carry out extensive field trials of an assessment instrument constructed according to Duplex Design specifications. Administrative and field services for the studies were provided by National Opinion Research Center (NORC).

8.3 The Illinois and California field trials of the Duplex Design

The goal of these trials was to test the implementation of a Duplex Design for 8th-grade mathematics and to evaluate the quality of the resulting data. Items for the design were drawn from the Illinois and California assessment programs and from the Second International Mathematics Study. Trials of the design were carried out in stratified samples of schools in Illinois and California. The Duplex instrument was administered by local school personnel under the supervision of NORC interviewers who were resident in the two states. The interviewers contacted the schools by telephone to arrange the testing and made one visit to instruct the classroom teachers in the method of administering the tests. All the materials for the testing were sent to and returned from the schools via United Parcel Service. The logistics of carrying out the studies were not difficult on a pilot basis and could be even more effectively implemented in a continuing program.

Our greatest concern in these studies was the student-level scoring discussed in Chapter 5. The success of the school-level scoring, discussed in Chapter 6, was never much in doubt; this aspect of the data analysis was not materially different from that used in the California Assessment Program since 1980. The critical question was whether procedures that were intended to yield student scores suitable for guidance and certification were workable and efficient. In particular, we hoped to demonstrate that two-stage testing could reduce, by at least a factor of two, the time required to administer the test in comparison with a conventional one-stage test. Without effective two-stage testing, the saving of classroom time that makes matrix-sampling assessment attractive would be lost, and from the teacher's point of view, the procedure would seem more like time-consuming achievement testing than like educational assessment.

For the present studies, we adopted the two-stage procedure described in Chapter 4. A 15-item pretest was administered on a day previous to the second-stage testing; the teacher or an assistant scored the pretest and inserted an answer sheet, which also had space for the second-stage responses, inside the front cover of a second-stage booklet of a suitable difficulty level according to the pretest scores. (The covers of these booklets were trimmed so that the students' names could be seen.) There were three levels of difficulty of the second-stage booklets—Easy, Medium, and Difficult. On the day of the second-stage testing, the booklets were returned to the students and testing proceeded in the usual way.

From the internal evidence of the data, and the responses of teachers to the NORC interviewers and on a debriefing questionnaire, this new approach to two-stage testing worked smoothly in all of the total of 64 schools participating in the Illinois and California studies. We consider our experience in these studies to have established that two-stage testing is quite feasible for the locally administered testing on which state assessment programs depend.

The remaining point to be established was that the desired two-fold gain in efficiency could be obtained by means of a 15-item pretest and a three-part second-stage test. All adaptive testing procedures, including two-stage testing, require some form of scaling of the item responses, rather than number-right scores, for expressing student attainment levels. In our case, we used IRT procedures based on the Bock & Aitkin (1981) marginal maximum likelihood methods of estimating item parameters. For the scoring of the students, we use Bayes estimation methods that are superior to all other methods in their overall precision in the population.

The present studies are the first applications of these types of IRT methods to two-stage testing. Frederic Lord (1980) had examined the theoretical advantages of two-stage testing, but he assumed joint maximum likelihood estimation of item parameters and maximum likelihood estimation of student scores. Our marginal maximum likelihood methods make better use of all information available to the analysis, however, and are more robust. The evaluation of the procedures discussed in Chapter 5 revealed that, in the overall reliability of the second-stage scores, the information curves for the test forms, and the relative efficiency of two-stage versus one-stage testing, our goals for two-stage testing were met or exceeded. Reliabilities for the student diagnostic profile scores were generally in the 0.80's, and a reliability of 0.95 was indicated for the students' overall mathematics attainment scores. The information curves exhibited the broad coverage required of tests that are to be used over the wide range of attainment levels seen at given grade levels in a state population. The relative efficiency of the two-stage test exceeded two in the higher and lower ranges of the score distributions where dependable measurement is most critical and most difficult.

We are conscious that adaptive testing in general, and two-stage testing in particular, has seen very little use in either educational assessment or traditional achievement testing. From the results of the present study, the conclusion seems clear that, not only is two-stage testing feasible, but in not taking advantages of these new testing technologies, half of the classroom time devoted to external testing is currently being wasted.

This judgment is not meant to imply, however, that a Duplex Design using two-stage testing can be administered in the same time required for a conventional group-level matrix-sampling instrument. In the present application of two-stage testing, we required classroom time for the first-stage test and student-background information, on a day prior to the main, or second-stage testing. In addition, we use an entire class period for the second-stage test in a single subject matter, whereas a pure matrix-sampling design would cover perhaps four different subject matters in the same amount of time. If satisfied with somewhat less individual diagnostic information in each subject-matter area, however, one might adopt a Duplex instrument that covers two subject-matter areas in one class period. But to attain some basic individual-level scores in each of four subject-matter areas would hardly be possible in less than two 50-minute class periods and perhaps another 30 minutes for the earlier first-stage test and questionnaire.

On the other hand, if *one-stage* testing were employed, the individual forms of the Duplex instrument would be equivalent to conventional achievement tests and would require the same amount of administration time. In that case, a marginal benefit from the design would still be provided by the greater generalizability and by the detailed school-level information that arises from the matrix sampling of the multiple Duplex forms.

8.4 The place of the Duplex Design in a comprehensive educational assessment program

Although the Duplex Design provides a more complete information system than either matrix sampling or achievement testing alone, it should not be considered a complete assessment system. Because it requires many items, economic considerations limit it largely to the multiple-choice items that can be mechanically scored. Many important objectives of education simply cannot be assessed by these types of items, however. In many cases, they will be accessible to large-scale assessment only if the resources are available to permit human readers to judge and score more informative types of assessment exercise. A number of states, including California, are already administering and scoring direct writing exercises. Procedures for developing the prompts to which the students write brief passages, and for organizing reading teams that subjectively rate the productions have been worked out and implemented. Statistical methods for scaling the ratings and maintaining their consistency form year-to-year have also been devised and successfully applied.

Apart from economic constraints (due to the costs of having extended responses read and rated), there should be no obstacles to extending the types of procedures used in direct writing assessments to similar evaluations of student productivity. These include the traditional "open-ended" questions that are composed by teachers and presented to students as part of the instructional process. From the steps that the student follows and records while answering the question, information about the understanding of the concepts involved or skill in the procedures required is revealed in a way that is almost impossible to duplicate with multiple-choice items. Although open-ended questions are perhaps more difficult to score than essays and written exposition, effective scoring protocols can be devised if there is sufficient provision for pretesting. Examples of this approach applied to mathematics may be found in the publication, *A question of thinking: a first look at students' performance on open-ended questions in mathematics*, prepared for the California State Department of Education.

A still more elaborate extension of the direct assessment of student production is the "practical" examination. This type of testing, in which the student manipulates equipment and material, has always been part of the laboratory sciences, creative arts, and the manual arts. Although testing these skills requires physical performance and some arrangement for rating the results, there is no other way to obtain the desired information—certainly not by means of paper-and-pencil instruments. For the most part, schools have the space and equipment for such testing; the impediment to their use in assessment is logistical and procedural, rather than lack of facilities. With enough resources and effort, however, practical testing could be brought into a program that also included open-ended exercises, and direct-writing tasks.

Even in such a comprehensive system, there undoubtedly will still be a need for objective testing based on multiple-choice items and the Duplex Design. Although such testing does not include the complete response repertoire of the student, it remains the most efficient way to cover a wide range of contents and certain types of procedural and reasoning skills. For this reason, objective testing is likely to continue to be an important part of large-scale assessment. Its further development through improved item content, instrument design, and scoring procedures should therefore not be neglected. The studies reported here are intended as a contribution to such development.

REFERENCES

Begle, E. G. & Wilson, J. W. (1970). Evaluation of mathematics programs. In E. G. Begle (Ed.) *Mathematics education. Sixty-ninth yearbook of the National Society for the Study of Education.* Chicago: University of Chicago Press, pp. 367–404.

Bell, M. (1972). *Mathematics uses and models in our everyday world.* Stanford, CA: School Math Study Group.

Bloom, B. S. (Ed.) (1956). *Taxonomy of educational objectives.* New York: McKay.

Bock, R. D. (1966). Contributions of multivariate statistical methods to educational psychology. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology.* Chicago: Rand-McNally.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29–51.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York: McGraw-Hill. (1985 reprint, Scientific Software, Inc., Mooresville, IN.)

Bock, R. D. (Ed.) (1989). *Multilevel analysis of educational data.* New York: Academic Press.

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

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement*, **12**, **(3)**, 261–280.

Bock, R. D., & Mislevy, R. J. (1981). An item response model for matrix-sampling data: The California Grade Three Assessment. In D. Carlson (Ed.), *Testing in the states: Beyond accountability*, pp. 65–90. San Francisco: Jossey-Bass.

Bock, R. D., & Mislevy, R. J. (1988). Comprehensive educational assessment for the states: the duplex design. *Educational Evaluation and Policy Analysis*, **10**, 89–105.

Bock, R. D., Mislevy, R. J., & Woodson, C. E. (1982). The next stage in educational assessment. *Educational Researcher*, **11**, 4–11, 16.

Bock, R. D., Muraki, E., & Pfiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational*

Measurement, **25**, 275–285.

Bock, R. D., & Zimowski, M. F. (1989). Sex differences in the mental processing of words and images. (submitted for publication.)

Brophy, J. (1987). Synthesis of research on strategies for motivating students to learn. *Educational Leadership*, **45**, 40–48.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average. *Educational Measurement Issues and Practice*, **7**, 2, 5–9.

Cohen, M. (1988). Designing state measurement systems. *Phi Delta Kappan*, **69**, 583–588.

Finn, J. D. (1974). *A general model for multivariate analysis.* New York: Holt, Rinehart, and Winston.

Finn, J. D. & Bock, R. D. (1989). *PC-MULTIVARIANCE: Univariate and multivariate analysis of variance, covariance, regression and repeated measures.* Mooresville, IN: Scientific Software, Inc.

Gadanidis, F. J. (1988). Problem solving: the third dimension in mathematics teaching. *Mathematics Teacher*, **81**, 16–21.

Gibbons, R. D., Bock, R. D., & Hedeker, D. R. (1989). Conditional dependence: Final report. Office of Naval Research report, contract #N00014-85-K-0586.

Goldstein, H. (1983). Measuring change in educational attainment over time. *Journal of Educational Measurement*, **20**, 369–377.

Jöreskog, K. G. & Sörbom, D. (1989). *PC-LISREL 7.* Mooresville, IN: Scientific Software, Inc.

Kilpatrick, J. & Wirszup, I., (Eds.) (1969). Problem solving in arithmetic and algebra. *Soviet studies in the psychology of learning and teaching mathematics*, **Vol. 3**. Stanford, CA: The School Mathematics Study Group.

Kilpatrick, J. & Wirszup, I., (Eds.) (1970). Problem solving in geometry. *Soviet studies in the psychology of learning and teaching mathematics*, **Vol. 4**. Stanford, CA: The School Mathematics Study Group.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Statistical Society of Edinburgh*, **61-A**, 273–287.

Lester, F. K. & Garofalo, J. (Eds.) (1982). *Mathematical problem solving: Issues in research.* Philadelphia: Franklin Institute Press.

Lord, F. M. (1962). Estimating norms by item sampling. *Educational and Psychological Measurement*, **22**, 259–267.

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

Lord, F. M., & Novick. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics.* **8**, 271–288.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika.* **49**, 359–381.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Maximum likelihood item analysis and test scoring—logistic models.* Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J., & Bock, R. D. (1989). *BILOG 3.0: Maximum likelihood item analysis and test scoring—logistic models.* Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J., & Bock, R. D. (1989). A hierarchical item-response model for educational testing. In R. D. Bock (Ed.), *Multilevel analysis of educational data*, pp. 57–74. New York: Academic Press.

Muraki, E., Mislevy, R. J., & Bock, R. D. (1987). *BIMAIN: a multiple-group item analysis and test maintenance program.* Mooresville, IN: Scientific Software, Inc.

Pandey, Tej (1989). Development of innovative questions for large-scale assessment. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics.* Washington, DC: American Association for the Advancement of Science (in preparation).

Pollak, H. O. (1970). Applications of mathematics. In E. G. Begle (Ed.) *Mathematics education. Sixty-ninth yearbook of the National Society for the Study of Education.* Chicago: University of Chicago Press, pp. 311–334.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan*, **68**, 679– 682.

Raudenbush, S. W., & Bryk, A. S. (1989). Quantitative models for estimating teacher and school effectiveness. In R. D. Bock (Ed.), *Multilevel analysis in educational research*, pp. 205–234. New York: Academic Press.

Resnick, L. B., & Ford, W. W. (1981). *The psychology of mathematics for* *instruction.* Hillsdale, (NJ): Erlbaum.

Roeber, E. J. (1988). *Survey of large-scale assessment programs.* Washington: Association of State Assessment Programs.

Schilling, S., & Bock, R. D. (1989). Expressing state assessment results on nationally normed scales. Paper presented at the annual meeting of the Education Commission of the States, Boulder, Colorado.

Schoenfeld, A. H. (1985). *Mathematical problem solving.* New York: Academic Press.

Sebring, P. A., & Baruch, R. F. (1983). How is NAEP used? Results of an exploratory study. *Educational Measurement: Issues and Practices.* **2**, 16–20.

Serötnik, K., & Wellington, R. (1977). Incidence sampling: An integrated theory for "matrix sampling". *Journal of Educational Measurement,* **14**, 343–399.

Shepard L. A., & Kreitzer, A. E. (1987). The Texas teacher test. *Educational Researcher,* **16**, 22–31.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika,* **42**, 425–440.

Skemp, R. R. (1987). *The psychology of learning mathematics.* Hillsdale, NJ: Erlbaum.

Stodolsky, S. (1988). *The subject matters: classroom activity in math and social studies.* Chicago: University of Chicago Press.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika,* **51**, 567–577.

Thissen, D., Steinberg, L., & Mooney, J. (1987). Trace lines for testlets: A use of multiple-category response models. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Tyler, R. W. (1956). *Basic principles of curriculum and instruction.* Chicago: University of Chicago Press.

Usiskin, Z., & Bell, M. (1983). *Applying arithmetic: A handbook of applications of arithmetic.* Chicago: Department of Education, University of Chicago.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement,* **24**, 185–201.

Wood, R. (1968). Objectives in the teaching of mathematics. *Educational Research*, **10**, 83–98.

Wood, R. (1971). Computerized adaptive sequential testing. Unpublished Dissertation, Department of Education, University of Chicago.