COMPREHENSIVE EDUCATIONAL ASSESSMENT FOR THE STATES: THE DUPLEX DESIGN

R. Darrell Bock
and
Robert Mislevy

CSE Report No. 262

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

1986

The project presented, or reported herein, was performed pursuant to a Grant from the Office of Educational Research and Improvement/Department of Education (OERI/ED). However, the opinions expressed herein do not necessarily reflect the position or policy of the OERI/ED and no official endorsement by the OERI/ED should be inferred.

Summary

State testing programs often attempt to provide annual information for use in student guidance and qualification, school and program evaluation, and for broad policy decisions. For these purposes, the programs have had to carry out several independent testing efforts based on different test instruments. In some cases, they have concurrently operated minimum competency testing, local achievement testing, and sampling assessment of curricular objectives. Because much the same content is covered in these tests, considerable duplication of costs and classroom time is necessarily involved.

With the development of a new type of assessment instrument, called a "duplex design", these several functions of state testing programs can be served in a single test administration requiring no more time and resources than conventional student achievement testing. Employing a multiple-form instrument similar to that used in sampling assessment, the duplex design yields profiles of individual student achievement in main content areas, while producing from the same item responses a detailed evaluation of curricular objectives at the school, district, county and state levels. Thus, when used in a state-wide census of student attainment, the duplex design serves in one comprehensive assessment the needs of diverse parties to public education for information on student achievement, school performance, and system-wide progress in attaining educational goals.

An example of a duplex design for eighth-grade mathematics illustrates the construction of the assessment instrument. Various applications of such instruments, including linking of results from distinct duplex designs for purposes of between-state comparisons, show the potential of this type of testing program. A technical appendix outlines the statistical model by which attainment scores for individual students are estimated on the same scale as those measuring detailed curricular objectives at the school level.

COMPREHENSIVE EDUCATIONAL ASSESSMENT FOR THE STATES: THE DUPLEX DESIGN[1]

R. Darrell Bock
University of Chicago
and
Robert J. Mislevy
Educational Testing Service

According to a 1985 survey, 47 of the 50 states mandate some form of statewide testing of student attainment (Winfield, 1986). These testing programs vary widely in design: some employ traditional every-pupil achievement testing, others are limited to minimum competency testing, still others make use of matrix sampled assessment at benchmark grade levels.

The most widespread program is minimum competency testing: 23 states have centrally directed programs, and another 16 allow local options of test content and administration; in 23 of these 39 states, satisfactory performance on the test is a requirement for high school graduation. Standards for passing are set variously by state legislatures, state boards of education, and local education authorities.

Many states have multiple programs, usually some combination of outcome assessment and individual achievement testing. States that have achievement measurement or minimum competency programs test every pupil at selected grade levels, but some of those using matrix sampled assessment test in a sample of schools. Others, such as California, use matrix sampling methods, but test in all schools.

California is a prime example of a multiple-program state: The California Assessment Program provides curriculum-oriented evaluations of school outcomes; local school systems are required to conduct their own minimum competency testing; and

[1]We are indebted to Linda Winfield, Leigh Burstein, David Wiley, Zalman Usiskin, Tej Pandy, and Mervin Brennan for valuable suggestions. Preparation of this paper was supported in part by the Center for Student Testing, Evaluation and Standards, School of Education, UCLA, and in part by a grant from the Spencer Foundation.

data from the National Assessment of Educational Progress (NAEP) are available in California for purposes of comparison with national results.

States that have no centrally directed program may nevertheless require the districts to conduct periodic achievement testing. In Iowa, all districts test annually and, in fact, all use the same test. Finally, end-of-high-school tests in specialized subject matter areas are administered to selected students in some states (New York State Regents Examination, California Golden State examination). Winfield (1986) and Burstein, et al. (1985) give detailed accounts of existing and projected state testing programs.

Considering that the information needed to assess educational productivity must be much the same in all states, the variety of these programs is at first glance surprising. Closer examination reveals, however, that they arise from different emphases on outcomes for which schools should be held responsible. Where the main concern is certification level of essential skills and knowledge, minimum competency testing is emphasized. Where the focus is on student attainment at all levels, especially when student guidance is involved, a commercial achievement testing program is usually relied upon. Where progress toward detailed curricular objectives is monitored, a matrix-sampling assessment program is the only practical approach. To the extent that mandated testing is committed to these disparate goals, the multiplicity of the existing state programs, with limited comparability of the resulting data, would seem to be inevitable.

We will argue, however, that with a suitable measurement design, a single, comprehensive assessment program can serve all of these purposes. We base this conclusion on an analysis of the information needs of the main users of educational test results within the states. The design we propose meets their needs directly and efficiently. In particular, it provides measures of achievement suitable for certifying attainment, for counseling students and parents, and for monitoring the effectiveness of schools and school districts. At the same time, it offers the detail and precision necessary for the evaluation of instructional methods and materials, and for basic educational research. Moreover, it performs these functions in a cost-effective manner.

The first part of this paper is an account of the thinking that led to what we call the "duplex design" for educational assessment. We then describe the design and its properties, display an example, discuss applications, and, in a technical appendix, formulate a statistical model for analyzing the resulting data. We also suggest how results from independent state assessments based on the duplex design can be referred to a common scale to allow comparisons among states.

1. The one best assessment design: an analysis of the information needs of various parties to public education

Anyone concerned with the conduct of education is conscious of the need for regular appraisals of student progress. Without such information, there can be no objective basis for guiding the student, for planning instruction, for evaluating schools, school systems and programs, or for correcting deficiencies or rewarding progress. It is not as well understood, however, that different forms of information about educational outcomes are required in these different applications. For example, test scores used individually in student guidance must be much more reliable than those that are averaged together to evaluate schools or programs. In contrast, the achievement profiles used in guidance seldom deliver more than six or eight accurate part scores, whereas evaluation of schools and programs may require 30 to 40 distinct measures if many objectives of the curriculum are to be examined. At the state level the information requirements of educational policy decisions are less demanding: relatively low precision scores can be aggregated into accurate group statistics, and indices in only a limited number of basic attainment areas need to be considered.

Although the information in all of these types of applications ultimately comes from the responses of the students, the different intended uses of the results influence the measurement procedures by which the data are collected and the statistical methods by which they are analyzed. Clearly, the first step in formulating the design and analysis must be a survey of the anticipated uses of the results. We begin this survey by identifying the potential *users* of the information, the decisions they will base on that information, and the characteristics of the information required in those decisions. We delineate seven categories of such users.

Teachers, school counselors, parents, and the student. Standardized individual achievement tests, independent of particular teachers or courses, are widely used as aids to informed and fair decisions on student advancement and placement. In addition to this every-pupil "summative" testing, teachers or counselors may, of course, also test particular students for "formative" or diagnostic purposes, but this type of testing is not within the purview of mandated programs.

Ethical considerations require that achievement test results be shared with the student and the student's parents. In this role, the tests must have three important characteristics: 1) they must cover content that is relevant to the course work for which the student is responsible; 2) they must be sufficiently reliable that scores on alternative forms of the same test will, with high probability, lead to the same recommendations on individual advancement or placement; 3) the results must be presented in a form readily understandable to the parties involved.

Typically, content coverage is assured by specification of domains defined by a taxonomy of subject-matter topics and objectives. Items then are written to conform to each of the classes of the domain specification. The validity of the classification for particular items may be checked empirically by inspection of the item-by-test score correlations, or by factor analyzing intercorrelations among items in a given content area. For the most part, writing appropriate items is reasonably straightforward once the domain specifications have been agreed upon.

To construct from such items a number of test forms that will produce consistent differential measurement of students is, however, a more difficult task. The problem is that the consequential decisions about students are made at all levels of the grade-level distribution: low ranking students may be kept back or sent to remedial programs; high ranking students may be put ahead of their grade or assigned to honors programs; students in the middle range may be assigned to tracked classrooms differentially. To be accurate over this range, an achievement test must have a sufficient number of items to measure accurately at difficulty levels throughout the expected score distribution. To span this wide a range, an individual achievement test must be rather long.

The large number of items required for accurate scores is an inherent limitation of such tests. The time available restricts the number of proficiencies that can be tested to a relatively small number. A test that reliably estimates achievement in six areas, for example, may require three to four hours of testing time. One of the problems we will consider in designing the comprehensive assessment is how to reduce the time required for dependable measurement of individual student achievement. New methods of adaptive testing, described below, make such savings possible.

Concerning the communication of achievement test results to teachers, parents, and students, a relevant observation is the normative nature of guidance-oriented use of test information. Teachers rarely make decisions about the student on an absolute basis; they can single out for special treatment only those students who deviate from the local standard. Because only rank-order information is required for such decisions, any form of reporting that indicates the student's standing in a reference group is suitable. This interpretation of test results is called "norm referenced". It is readily understood in the context of student guidance. Test publishers therefore strive to maintain accurate and up-to-date percentile norms from a relevant population.

If accurate norms extend over grade levels in which students are tested annually, the scores enable the teacher to see the trend in a given student's progress from changes of his or her position in the score distribution over time. Alternatively, statistical methods can be used to express the scores at different grade levels on a common scale of measurement. The use of Thurstone's absolute scaling method to define a "grade-equivalent" scale is an example. Although this type of scale seems easy to interpret, it has the disadvantage that the differences between grades from 1 to 12 do not represent equal steps in the development of attainment. Because the standard deviation of scores within grades is larger for higher grades, it is much more serious for a third grader, for example, to be a year behind grade level than it is for an eighth grader. Indeed, at the eighth grade level, a grade equivalent may be less than one standard deviation of the within-grade distribution. Thus, it is quite possible for 20 percent of eighth grade students to be one year behind grade level owing to normal individual differences within grade. The wide range of within-grade variation when expressed on the grade equivalent scale also leads to seemingly anomalous situations where, for example, a superior fourth grade student is reading at the 8th grade level. Where normative reporting is required for the duplex design, we will adopt grade-normed standard scores rather than grade equivalents as the reporting medium.

Regardless of the statistic used to express relative standing, its precision will be limited by the number of items that can be included in the achievement measures in the available testing time. For this reason, individual achievement results should always be reported with a standard error or as a confidence interval. If the latter is shown graphically, even a person unfamiliar with statistical concepts has some appreciation of the uncertainty present in the test scores. The statistical methods we propose in the appendix provide accurate standard errors for these purposes.

To summarize the needs of teachers, counselors, parents, and students for information about individual student progress, we can say that they require measures that are precise enough to be depended upon in guiding students, are reported in normative terms, convey the uncertainty of the scores, and have as much diagnostic detail as is possible within the time constraints of an external testing program. These needs are served by well-developed educational testing technology based on the standardized achievement test—typically a machine scorable multiple choice test, published in several parallel forms, and focused on subject matter appropriate for specific grade levels or courses.

Designers of curricula and planners of instruction. A quite different kind of information about attainment is required by persons designing curricula or developing instructional methods and materials for the classroom. In these applications, it is not the individual student that is to be evaluated, but the overall performance of students taught under different conditions. Although the classroom teacher has an interest in the outcome of such evaluations, it is primarily the school department head and principal, the professional curriculum specialist, and the textbook and workbook writer who will make direct use of these results.

Persons involved in curriculum research need a much more detailed description of student attainment than is available in traditional achievement testing. The problem is that measures of broad content areas produced by achievement tests are insensitive to differential curricular effects. Although this fact has not been emphasized in the evaluation literature, it has been amply demonstrated in empirical studies of alternative curricula. Walker & Schafforzich (1973), in a lengthy review of research on science and mathematics curricula from 1956 to 1972, found that any given curriculum tends to be superior to others only in respect to material that is distinctive to it. Where the content and presentation are common among curricula, all perform equally well; thus, the differential outcomes are seen in contrasting score profiles, not in overall performance. An important corollary of this finding is that the tests employed in such comparisons must be sufficiently detailed to measure separate outcomes for distinctive parts of the curricula. An instrument used to evaluate "new" math and traditional math, for example, would have to produce reliable scopes for both of these types of content. Conventional achievement tests typically do not deliver scores at this level of detail.

By the same token, instructional planners need to examine student performance in the units of content that can be manipulated in instruction. To write lesson plans for mathematics, for example, the instructors need to know the specific units–computation, number systems, problem-solving, applications, etc., that need attention. These units are almost always tested formatively, but time restrictions prevent their separate evaluation during external achievement testing.

To be useful to this constituency, an evaluation instrument must distinguish perhaps 20 to 40 curricular objectives at a given grade level. Because it is not possible to test this many topics with the same precision that is demanded of individual measurement, a quite different approach is required. The key to this approach lies in the fact that individual measurement is not necessary in program evaluation; only the average performance of classrooms or other experimental units need be measured. If the number of students in these groups is sufficiently large, good precision in estimating program effects can be obtained without the use of long tests. The generalizability of the group mean scores is the important consideration, not the reliability of scores for individual students.

It has been known for some time that to obtain adequate generalizability in estimating program effects, evaluation should not be based on the traditional achievement test, but on an instrument in which each student responds to only a few items sampled from each of numerous content elements, while different students respond to different samples of items. This approach assures good generalizability of the group mean for each element with minimal demands on testing time. It is the basis for the multiple matrix sampling designs used in the National Assessment of Educational Progress and in numerous state testing programs. In these designs, the test instrument is constructed in many forms, 15 to 30, or sometimes more, with a small number of items assigned randomly to each form from the pool representing each curricular objective or element. Lord (1980) has shown that the most efficient matrix sample is one in which each student in the group is assigned one distinct item from each element. In that case, the number of curricular objectives that can be assessed in one form is then equal to the number of items that the student can respond to during the testing period, usually 30 to 40. This number is quite adequate for a highly detailed curricular evaluation.

The scoring of matrix sampled instruments is also different from that of achievement tests. In the original formulation of matrix sampling (see Lord, 1962), the scores are not presented in any normative form, but simply as average percent correct for each content element. Classrooms, groupings of students, instructional programs, schools, and other aggregations are then compared with respect to the strengths and weaknesses revealed in the profile of average percent correct scores over detailed curricular elements. Since these elements usually correspond to units or topics of instruction, definite recommendations about teaching practices or emphasis can be made from such results.

More recently, Bock, Mislevy and Woodson (1981) have shown how matrix-sample data can also be analyzed and scored by use of scaling techniques based on item response theory (IRT). According to this theory, the probability that a student will respond correctly to a given test item is a function of the student's location on the proficiency dimension and of properties of the item, such as its difficulty and validity. The properties of each of the items in a test can be estimated from large samples of responses and used to estimate a "scale score" for the student indicating his or her proficiency level.

Average percent correct scoring and IRT scale scoring both retain the detail necessary for curricular evaluation and instructional planning, but scale scores have the advantage of remaining comparable as items are added to or retired from the instrument from time to time. This consistency of interpretation as the item content is updated is essential if educational progress is to be followed over long periods of time. Recently developed IRT test maintenance systems (Bock & Muraki, 1986) even provide for the detection and correction of drift in the relative difficulties of items that may occur over time.

Local school system managers, officers, and boards. In making decisions on personnel, resource allocation, and policy, school officials must be able to support their actions with data on educational outcomes in the schools for which they are responsible. In addition to such operational statistics as number of students in school, number of hours of schooling, teacher/student ratio, etc., they need measures of outcomes in the relevant subject matter areas at a number of grade levels. The detail required depends somewhat on the style of administration or oversight of the persons involved. Superintendents and boards that have considerable experience with education and instruction probably will be interested in more detail than is available from achievement testing, although perhaps not to the same extent as the curriculum specialist. They will not, however, be interested in a level of score reporting below that of the classroom or school. Because their concern is with group-level rather than individual outcomes, they can make profitable use of the matrix sampling methods of program evaluation. The only difference is that classrooms or schools rather than programs are being evaluated, a distinction that is conveyed by describing the activity as "assessment" rather than "evaluation".

Assessment procedures based on matrix sampling designs have the advantage of providing a detailed profile of aggregate outcomes, without intruding excessively on classroom time. Equally advantageous, however, is their resistance to effects of "teaching to the test". Because there are so many items in the forms that make up assessment instruments, it is difficult for a teacher to discuss enough of the items to have any great effect on the school outcome. Indeed, an attempt to teach a majority of the items would be virtually equivalent to teaching the subject matter of the course. In addition, if scale scoring is used, a proportion of items can be replaced periodically to protect further the integrity of the test.

Achievement tests, in contrast, typically exist in only a few forms and are not always updated regularly. If school districts use the same achievement tests from year to year, the items tend to become known to the teachers, who may then consciously or unconsciously teach the specific information required to answer particular items. If so, the tests will tend to show year-to-year average gains that do not reflect increased general knowledge of the subject matter on the part of the student. The more pressure the teachers are under to improve student outcomes, the greater the probability that these teaching-to-the-test effects will appear.

Whether the information on student progress comes from achievement tests or assessment, it is important to school officials that the scores be reported on a scale with fixed origin and unit so that gains or losses in each subject matter area can be compared over a period of years. The sort of rank order information that is acceptable for comparing individual students is not suitable for monitoring the progress of schools and school systems. Average number correct scores in assessment results have this property, but they have the disadvantage of losing their comparability if some items are retired from or added to the content areas assessed. As Lord (1980) has discussed, IRT scoring of tests facilitates both the equating of test forms and the updating of item content within forms. This theory also allows accurate calculation of measurement error variances at all points on the scale. These error-variance estimates can in turn be used in obtaining efficient, weighted estimates when aggregating data to the school or district level, and in expressing results in the form of confidence intervals that convey uncertainty due to the sampling of both students and items. We discuss below these and other contributions of item response theory to educational assessment.

State departments of education. The activities of most state departments of education are sufficiently varied to benefit from all of the outcome measures described above. Departments that formulate curricula or set objectives need feedback from detailed assessment of curricular objectives. Most states employ for these activities professional specialists whose work depends critically on this type of information. At the same time, most departments of education are also concerned with the performance of schools as measured by numbers of students reaching or exceeding defined levels of achievement, whether minimal, ordinary, advanced, or outstanding. For these purposes, individual achievement measures in broad subject matter areas are required. For just this reason many states operate assessment programs simultaneously with conventional, in many cases commercial, achievement testing.

Some states have limited assessment programs based on sampling of schools and students within schools. If the state also has a policy of accountability of school districts for levels of student attainment, however, this type of sampling is not sufficient, and a complete census based on every-pupil testing has to be implemented. The effort can be well repayed: because the census provides accurate information at the level of the individual schools, results can be reported in a form that is interesting and informative locally, and schools with exceptional outcome patterns can be identified throughout the state. If the state makes space grants to improve average student performance, or rewards such performance financially, then a complete census is, of course, essential.

An additional problem with a sampling assessment is that the schools have no immediate payoff. Motivation for cooperation on the part of both staff and students is minimal, and levels of performance may suffer as a result. Apart from the lower cost of sampling assessment, there is little to recommend it over an every-pupil program.

The quality of information that state departments of education have at their disposal is also generally better when the test data take the form of original response records of the individual students. Although districts may have the capability of scoring tests and reporting summary statistics, the information can be analyzed more consistently and in more detail if primary rather than secondary data are available to the department.

State legislators and officials. At the state level, representatives not exclusively involved in education can attend only to rather general indices of educational outcomes. They cannot go into the detail that would interest the curriculum specialist, or even the more limited achievement profiles required for student counseling. Their concern is primarily with the main subject matter areas measured at a few benchmark grade levels, e.g., 4, 6, 8, 10 and 12. Often, year-to-year gains and losses are of more interest than absolute levels of attainment. The statistics necessary for these general summaries of educational progress can readily be obtained by aggregating the more detailed assessment figures at the school or district level. The precision and generalizability of these statistics will be so high that the confidence intervals required at lower levels of aggregation will seldom be necessary, although they can be calculated if required. If reported in the form of scale scores, the results will remain comparable over relatively long periods of time, and long-run changes in the average performance of students in the state can be traced.

Surveying information needs at the state level, one also foresees a demand for broader interpretations of educational outcomes, with efforts to explain findings and attribute causal relationships. To aid such interpretation, the Council of Chief State School Officers has encouraged efforts to compare student attainment among states (Selden, 1986). By taking into account the background and composition of the student population as well as the resources of the school systems, the states will then be better able judge the effectiveness of their curricula and schools. The comprehensive assessment system must therefore provide for eventual conversion of state results to common scales that permit between-state comparisons, at least with the degree of detail that is typical of the state level-reports. We discuss this problem in section 7.

Similarly, by examining annual assessment data, state officials may be able to infer the impact of current social trends on student performance (e.g., television viewing or microcomputer use). They may then be able to anticipate educational problems that will eventually influence public policy or legislation. Long-term stability and consistency of a state's assessment program and procedures are essential to such inferences.

The media and the public. Communicating data on school productivity in a form accessible to the general public is a challenging task for the educational evaluator.

The key to success is making the findings understandable to the journalists who must report such information in the newspapers and on radio and television. Past experience indicates that reporters have difficulty understanding the arbitrary scales in which attainment data have to be expressed. Reporting of average percent correct for a content area, which provides only relative information and varies in level from one content area to another, is especially troublesome. The writer's audience has to keep in mind that the scales are not comparable. A better practice is to employ scale scoring, defining a scale with a common origin and unit for all subject matter areas and employing it uniformly until its characteristics become well-known. Comparisons between schools or groups of students can then be expressed in familiar numbers, and year-to-year gains or declines in student performance can be followed in units that have a widely understood meaning. Certain achievement scales, such as that used to report Scholastic Aptitude Test (SAT) scores, have achieved this status.

An even more comprehensible form of reporting, however, is to state the percent of students who fall above or below certain thresholds on the attainment scale. If these points correspond to administrative cutting points (e.g., for graduation, special honors, admission to college, etc.) their practical implication is entirely clear. If these objective criteria do not exist, the item content typical of selected score levels can be exhibited to convey the nature of the tasks that students at these levels can typically perform. The NAEP reading scale, for example, is characterized for reporting purposes by displays of items that students at the 150, 200, 250, 300 and 350 points on the scale have an 80 percent chance of answering correctly.

Another possibility is to take a normative approach and designate certain arbitrary percentile points in the population of students. The 25, 50 and 75 percent points, for example, might be referred to as the "basic," "ordinary," and "advanced" mastery levels. In this connection, however, it must be mentioned that achievement testing and assessment are quite different when it comes to estimating the percent of students above a specified performance threshold. In achievement data, it is a simple matter to obtain these percentages by enumerating students whose individual scores fall in the defined intervals. But from matrix sampled assessment data, individual scores are not available, and the percent of students above some point on the scale of the group means can be estimated only if the distribution of proficiencies within the group can be described. Up to now, the information necessary to estimate these within-group distributions has not been part of assessment results; it has had to come separately from conventional achievement tests rather than matrix sampled assessment designs. One of the main strengths of the duplex design is that the proportions of students exceeding specified mastery levels can be estimated in the same manner as in achievement testing. This enables percents of students at specified levels to be estimated directly. Another excellent method of reporting, based on so-called "criterion-referenced" test scores is described below (section 6).

Educational research specialists. A constituency independent of school systems, yet having an interest in the information generated by state testing programs, consists of academic and professional research workers engaged in study of education and the schools. In principle, they can use information from either achievement testing or assessment. But like the curriculum specialists, they are also often interested in detailed areas of attainment, not just the broad skill areas measured by individual achievement tests. Thus, the data from assessment programs may be more relevant to them than traditional test scores. Assessment data will also typically have higher generalizability indices, and thus clearer relationships with other variables.

The latter fact is demonstrated by the effects on estimated correlations of increasing numbers of students sampled and items sampled (see Table 1). The data are reading score means in California schools measured in two successive years. Notice that the sizes of the correlations increase (the school means become more accurate) as the sizes of the samples of students increase from row 1 to row 3. Similarly, the correlations increase when student sample size remains fixed, but the numbers of items sampled increase from 40 in a single test form, to 128 in 16 forms, to 400 in 40 forms. This latter effect arises from the increased generalizability due to item sampling. It would be even more pronounced if different items were sampled each year. It would then maximally suppress the effects of item heterogeneity that attenuate relationships between student attainment and the background variables.

TABLE 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		
		40	128	400
Number of students	50	.59	.73	.79
sampled	100	.67	.78	.88
per grade	200	.76	.81	.93

Because most research workers depend for their data analysis upon standard computer packages that require scores for individual respondents, however, matrix sampled assessment data can present something of a dilemma. Only more advanced workers currently know from first principles how to use matrix sampled data directly (e.g., by empirical Bayes methods; see Mislevy, 1985). Until computer packages become available for analyzing item response data or scores that exist only at the group level, the data obtained from matrix sampling designs will not be convenient for secondary analysis. In this respect, the duplex design proposed in this paper has a marked advantage: it supports scoring of the same item response data at both the individual and the group level. Research workers can thus make use of either of these forms depending on their statistical expertise.

2. Summary of information uses

The uses of information on student attainment that are identified in the preceeding section can be classified in terms of the decision-making activities involved. The following five broad categories result.

Guidance: counseling, placement, promotion, and certification of individual students. Each requires accurate test scores in at least the main areas of proficiency and subject matter in the curriculum. Standardized achievement testing is a main source of this information.

Evaluation: choosing among competing curricula, instructional programs, or educational materials. These choices require information on the performance levels of groups of students pursuing alternative programs or using different materials. Matrix sampling assessment, making minimal demands on student testing time, provides this type of information at the group level, but scores for individual pupils are not available by this method.

Management: monitoring student attainment in programs, schools, and school systems. Managerial decisions can utilize measures of attainment at the classroom or school level. They need much the same level of detail as evaluation studies. Resistance to teaching-to-the test is vital in this use. This information need is better served by assessment methods than by individual student achievement testing.

Policy: judging the overall progress of an educational system, or its main components, for purposes of formulating legislation and allocating resources. Policy decisions can utilize statistics of attainment aggregated to the district or state level. They do not require the level of detail needed in program evaluation or school management. The required information can be obtained equally well by achievement testing or by assessment results summarized in broad areas of proficiencies or subject matter.

Research: secondary studies of the conditions and background variables that influence student attainment. Statistical methods in educational research typically depend upon accurate scores for individual students. The existence of widely used, well-defined scales for reporting results greatly facilitates such studies. Student achievement testing based on standardized measures has traditionally served this purpose.

Other conclusions follow from this survey of attainment information use. We have seen that, because of the limited time available for student testing, there is a trade-off between precision of individual measurement and breadth of content coverage. This trade-off is a major problem for the comprehensive assessment design. To serve all of the above purposes, the design must provide the precision of individual measurement required in guidance and research, while delivering information at the group level on detailed curricular objectives for purposes of evaluation and management.

We also noted the distinction between sampling assessments and those based on a total census of the state at selected grade levels. Whereas evaluation, policy, and research can make use of data from samples of schools and students, guidance and management require a total census of schools and pupils. Fortunately, the marginal cost of extending a sampling assessment to a census, once the systems of instrument development, test administration, scoring, analysis and reporting are in place, is relatively small in this age of computer data processing. Recently, these considerations have influenced a number of states that have relied on sampling assessment in the past to begin converting to an every-pupil testing program (e.g., Illinois, Missouri).

A conclusion applicable to all five of the above categories of use is that many benefits flow from a continuing program of measurement capable of producing dependable and comparable scores annually over periods of years or decades. Guidance can then be based on the developmental history of the student's attainment, rather than current status alone. Evaluation can look at effects of program interventions relative to a pre-intervention baseline within the program, instead of depending exclusively on comparisons between programs. Management can appraise the performance of schools relative to their own past performance, and not merely to that of other schools. Policy decisions can make use of indicators of student attainment that can take their place with other well-established indicators of social and economic change. Finally, research studies in education can take advantage of the typically stronger inference that is possible when growth and change can be examined directly within students, and intervention effects can be detected in sequential data from a single program, and not just in comparison between programs.

Gaining these benefits depends upon the design of a comprehensive assessment system that will serve all these functions effectively and efficiently. We devote the remainder of this paper to the solution of this problem.

3. Combining student achievement testing and assessment of curricular objectives

It should be clear from the preceding discussion that traditional individual achievement testing differs in important ways from the assessment of the progress of schools or programs in attaining specific curricular objectives. Up to now, these two types of educational measurement have been conducted in separate testing programs. Where both are employed in the same state, districts or local school systems are often responsible for achievement testing, while assessment of curricular objectives is the responsibility of the state program. In some states, assessment is conducted by sampling of schools and students, but in others a complete census of all students is carried out in the benchmark grades.

Although the instruments administered in these two types of testing differ, their item content at any grade level is much the same, and substantial duplication of cost, effort, and demand on classroom time is involved in obtaining the same information in different forms. We show in the present paper, however, that with a suitably designed assessment instrument, based on the duplex design, both of these forms of information can be obtained in a single test administration requiring no more classroom time than conventional achievement testing. The instrument we propose for this purpose has multiple stratified random test forms like those used in assessment, but the items are assigned to forms in such a way that a student's response to a particular form can be scored in broad skill areas, while responses over forms can be aggregated to provide scores for detailed curricular objectives at the school or other group level. An example of the layout of this type of instrument in the area of eighth grade mathematics is shown in Table 2.

TABLE 2

A GRADE 8 MATHEMATICS DUPLEX DESIGN

Content Categories		Proficiencies		
		a. Procedural Skills[2]	b. Factual Knowledge[3]	c. Higher Level Thinking[4]
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

[2]Calculating, rewriting, constructing, estimating, executing algorithms.

[3]Terms, definitions, concepts.

[4]Proof, reasoning, problem solving, real-world applications.

For this design, mathematics attainment is divided into three broad categories called "proficiencies". The mathematical content of the proficiencies is arranged in the content categories of the discipline as reflected in curricula and textbooks at this grade level. Scores for individual students can be calculated within forms for each of the proficiencies, an average score for the mathematics area as a whole can be obtained by averaging the three proficiencies. Scores for schools or other groups of students can be calculated for each of the 57 elements of the table for which suitable items are available. Depending on the item pool, not all of these elements may be included when the design is implemented. In a grade 8 mathematics design based on items from the California and Illinois Assessments, central categories 15 (irrationals), 23 (inequalities), and 52 (experiments and surveys) were not represented. If the scoring methods described in the appendix are used, the mean of the proficiency scores of pupils in a given school will equal the mean of the school-level content-element scores within that proficiency. Thus, the two types of information extracted from the duplex design are expressed on the same scale of measurement.

The items of the assessment instrument will constitute a complete, or almost complete, representation of the elements in Table 2, replicated randomly in, perhaps, 24 printed forms. The items in any given form will be chosen randomly from the pools representing each of the curricular elements.

In the administration of the instrument, these forms are distributed in rotation within classrooms. The fact that different pupils may be responding to different forms and items does not typically present any difficulty provided the covers of the forms are similar and any practice items presented on the first page of the forms are identical. This method of test administration has been used widely in assessment programs with good success. In particular, the experience of the California Assessment shows that, when expendable test forms are used, group testing with this type of instrument can be carried out as early as the third grade.

4. The contribution of modern item response theory (IRT)

To estimate comparable skill area scores for all students regardless of which test form they are assigned requires the use of modern IRT methods of test scoring. It is assumed that in a certain base year, the instrument has been administered to a probability sample of students at the appropriate grade level. The test items are then calibrated, preferably by the marginal maximum likelihood method (Bock and Aitkin, 1980), and the units of scale are chosen so that the mean and standard deviation in the population of students is the same for all skill areas. The resulting item parameters are then used to compute students' scores by maximum likelihood or Bayes methods, with each score accompanied by a standard error or posterior standard deviation. Scores computed in succeeding years with these item parameters have a constant origin and unit defined arbitrarily in the base year. They are thus suitable for measuring growth and change in the population or in subpopulations from that year onward. Because IRT methods are used, it is possible to add and retire items from the test without altering the base year definition of the scale. This can be done as part of the operational administration of the test and requires no additional calibration studies. As mentioned above, recent progress in item response theory makes it possible to account for effects of so called "item-parameter drift" while retaining the original scale definition (Bock and Muraki, 1986). This assemblage of measurement techniques, along with provisions for writing and evaluating new items, constitutes the item maintenance system that supports the comprehensive assessment program.

Scores for schools or other groups of students can be estimated by IRT methods using the models for group data described by Mislevy (1984). These methods provide scores for the curricular elements on the assumption that each pupil responds to one item from each element. The duplex design for the assessment instrument satisfies this condition. This type of scoring is especially easy to carry out because it uses, as statistics, the number of students who attempt each item within the classroom or school, and among those the number who respond correctly. Thus, the calculations require only a classroom or school summary file rather than the vastly larger file of individual item responses required for the scoring of students in the skill areas.

The group level scoring is also based on a calibration of the instrument during the base year. To distinguish these two types of calibrations and scoring, we refer to those for the student proficiencies as the "vertical" calibration, and those for the school, classroom, or other group scores on the separate curricular elements as the "horizontal" calibration. In the appendix to this paper, we present the mathematical model and estimation procedures by which the vertical and horizontal calibrations can be carried out so that the individual and group scores can be expressed on the same scale. Thus, both the student achievement (vertical) scores and the school assessment (horizontal) scores will give the same result when broad skill area and subject matter scores are aggregated for state-wide monitoring of educational progress.

5. Adaptive testing

With the aid of IRT scoring methods, it is possible to minimize testing time with the use of some form of adaptive test administration. Ideally, one would prefer individual, fully adaptive, computerized test administration in which each item presented to the examinee is most informative, given the provisional estimate of the examinee's proficiency at that point. But almost equal gains in efficiency can be obtained by group-administered, two-stage testing (Lord, 1980). In this form of testing, each student takes a short pre-test of general knowledge in the subject matter area. This pre-test is typically self scored by the student, who is then directed to a section of the main test where the level of difficulty is suitable for a student with a given pre-score. Bock, Sykes and Zimowski (1986) have reported a feasibility study of a form of two-stage testing especially suitable for the duplex design. In their instrument, the second-stage test consists of three replications of the item content represented by items of increasing difficulty spiraled in a single test form. At various points on the answer sheet are flags indicating where a student with a given pre-score should begin and end answering questions in this spiral. Each student who completes the items within the assigned block thus covers all of the item content at a level approximately suited to his or her general knowledge in that subject matter. Based on the results of the feasibility trial, these authors find that with items at typical levels of discriminating power, skill area scores based on 16 to 20 well positioned second-stage items will have a reliability of about .85 with respect to the population distribution of proficiency, and a reliability of about .97 for a subject matter area consisting of three skill areas. These levels of reliability would generally be considered high enough for the purposes of school achievement testing. To obtain them without two-stage testing would require at least twice as many items and essentially double the required testing time.

6. Criterion-referenced reporting

The most widely used method of expressing test scores in a standard form is to convert the score to a percentage point or standard deviation in some reference population. In achievement testing, this population is usually made up of students at the relevant grade level in the community, state, or nation. The standard score is thus defined, not in terms of the subject matter being tested, but as the standing of the student with respect to his or her peers.

This so-called "norm-referenced" method of interpreting test scores is useful in identifying students of special distinction (in either a positive or negative sense), but it does not specify concretely the degree of mastery of the subject matter the student's score represents. That even the lowest ranking student commands enough of the subject matter to apply it in some practical way or to go on to further studies is not necessarily conveyed by these scores. This sort of information for norm-referenced scores must be established in external validity studies of the test and documented separately in the test manual. Often, norm-referenced tests are extensively used without benefit of adequate validity studies or documentation.

It is therefore of considerable interest that, with the introduction of IRT methods of item analysis and test scoring, a method of interpreting tests scores directly in terms of the content of the test items has become available. With IRT calibration of test items, the probability of a student's correct response to a test item can be calculated from his or her proficiency scale score. This makes it possible to interpret the scale score in terms of the content of items for which the student has some arbitrary (typically 50 or 80 percent) probability of answering correctly; that is, the location of the items and the location of the student can be expressed on the same proficiency scale. The NAEP reading scale referred to in section 1 is an example.

This property of IRT methods permits us to define Linearly Ordered Content Domains (LOCD's) that represent the stages of content mastery that a student proceeds through as his or her proficiency increases. In the construction of an LOCD, it is not so much the location of the particular items that define the LOCD, but the *class* of items that can be written according to some specification. Using multilevel procedures for item parameter calibration, we can estimate the means and standard deviations of the locations of items sampled from the class specification. These statistics then characterize the level of attainment represented by various points on a proficiency scale. A great advantage of a criterion-referenced test is that it does not require a norming study. Used in any age group at any time of year, it still yields interpretable scores. Of course, normative information for a criterion-referenced test increases its usefulness, but, unlike the norm-referenced test, normative data are not essential for productive use of an LOCD.

Regrettably, only a few well-defined LOCD's exist at the present time. One of these is the basis of the Degrees of Reading Power (DRP) test published by the College Entrance Examination Board; it defines points on a reading proficiency scale in terms of classes of reading material (e.g., children's magazines, certain newspapers, types of textbooks, technical reports, etc.) that can be read with a specified degree of comprehension. A similar LOCD for spelling has been proposed by Wilson and Bock (1985), but not yet implemented in a published test. For proficiency in written expression, an LOCD could easily be constructed by publishing a collection of written passages ordered according to quality of expression by expert judges, but this has yet to be done. Constructing LOCD's in areas such as mathematics and science would perhaps be more difficult, but progress will undoubtedly be made as IRT methods of item analysis and test scoring come into wider use.

7. Linking state assessment results

The foregoing considerations underscore the potential of the duplex design, implemented by two-stage testing, for versatile, cost-effective meeting of state needs for information on student attainment. In the realm of policy formulation and research, the contribution of the duplex design to progress in education would be even greater if the assessment results of different states could be compared. The states could then more easily share findings and experience gained through monitoring of outcomes of their respective educational systems.

To make such comparisons with tests constructed independently by the states will require special provisions for establishing a correspondence between the scores on the separate tests. These provisions will impose two critical conditions on the state testing programs, viz., 1) sufficient similarity in the definition of content categories and proficiencies to provide a logical basis for equating the scoring scales, and 2) close enough correspondence of the conditions of testing—grade levels tested, time of testing, coverage of the student population, etc.—to justify comparisons without elaborate adjustments of the data.

In considering whether the first condition could be satisfied, it is important to realize that, although curricular specifications underlying the duplex designs would be independently arrived at by state communities, there are powerful influences toward uniformity. The committees that write curricula for the states necessarily depend upon the expert advice of teachers, university professors, and curriculum specialists in the relevant subject matter. Even when drawn from residents of the state, these experts almost always belong to nation-wide professional organizations and communicate through the same national publications. They inevitably tend, therefore, to reflect prevailing professional views on curricular objectives and subject-matter content. If the curricular committees include eminent educators from other states expressly for the purpose of avoiding a too-parochial approach to educational planning within the state, the tendency toward homogeneity is even greater.

The other great driving force for uniformity is, of course, the textbook publishers, who for economic reasons attempt to prepare teaching materials that have the widest possible range of use. Because, when preparing such materials, they all use much the same methods of surveying school practices, curricular conventions, and the composition of published achievement tests and other textbooks, the commercial publishers tend to produce very similar textbooks within subject-matter areas. Committees writing test item specifications tend, in turn, to rely heavily on analysis of textbooks to define content coverage. The process is more than a little circular, and a substantial amount of agreement among independently prepared tests necessarily results.

States will differ in the range of courses they require, especially in more peripheral topics, such as consumer education or personal health. But internal to a given subject-matter area, the underlying unity of the concepts can be seen in the similarity of the tests used to measure achievement tests in the area. The names chosen for the content categories and proficiencies may differ in the test specifications, but the similarity of the items assigned to them shows that the conceptual basis is the same. By formulating a nomenclature general enough to include the variance found in the local specifications of curricula or test designs, we can bring many of these test designs into correspondence.

Because the duplex designs are built up from elements defined by content categories and proficiencies, it is possible to move identifiably similar elements from different designs into a form that allows common scales to be constructed. Of course, not all of the elements in the various duplex designs will be in correspondence. But where they are, equating studies will relate the corresponding scale scores at the level of specific curricular objectives in the cells of the duplex design. Similarly, the partition of the paired cells into the proficiency measures in the columns of the design will result in a correspondence of scores at the level of the individual achievement measures.

Indeed, finding points of similarity in the duplex design specifications may be easier than negotiating the agreements required to meet the second of the above conditions, viz., common grade levels and times of testing. At present, there is considerable variation in the choice of the benchmark grade levels to be tested and on the time of testing. Almost all of the state testing programs include grade 8, but beyond that there is less concordance (Burstein, et al., 1985). Given that it is not essential to test at all grade levels, and excluding children in first and second grade as too young for paper-and-pencil testing, one might propose grades 4, 6, 8, 10, and 12 as logical choices for an intensive program, and grades 4, 8, and 12 for a less demanding effort. Unfortunately, a number of states test at grade 3 instead of 4, others test 11th grade in preference to 12th in order to avoid the many conflicting activities of the senior year in high school. Although it may be possible to develop a statistical model for changing attainment between grades that could be used to predict scores for grades not tested (for example, performance at grade 4 could be predicted from performance at grades 3 and 6), such adjustments would at best be inconvenient, and, at worst, would reduce our confidence in the validity of the cross-state comparisons.

Much the same is true of the time of testing. Although Burstein, et al. (1985), find that most states test in the spring, the few that do not would complicate the comparative analysis of state results. Though it would also be possible to make adjustment for gains during the year as a basis for comparison, the analysis would be simpler if the time of testing were approximately the same in all states, perhaps preferably during the Spring at all grade levels except 12, where Autumn testing, before students begin taking their college entrance examination tests, would ensure better motivation.

Inasmuch as grade level and time of testing are entirely arbitrary and have no educationally significant implications, this is one aspect of the state testing programs that could easily be aligned through the good offices of a national organization such as the Council of Chief State School Officers or the Education Commission for the States. Although in some cases the language of the legislation enabling the programs would have to be altered, the required changes are innocuous and should not meet any opposition.

If the above conditions can be met, the remaining problem will be how to express, on the same scale, results from those parts of the duplex designs that are common among states. In IRT based testing practice, there are three quite different methods available for equating tests:

1.) *Common item method.* If two or more tests purporting to measure the same quantity have a number of items in common, IRT methods can be employed to calibrate the entire set of test items on a common scale. This is possible even when the two tests are taken by examinees drawn from different populations. The common items link the results together during the calibration, and the scale scores estimated from the separate tests are then expressed commensurately and can be treated interchangeably for purposes of comparison.

2.) *Common population method.* If the tests to be equated are administered to large samples of examinees from the same population, the indeterminacy of the origin and unit of the IRT scale scores can be resolved by setting the mean and standard deviation equal in the two samples. Even though the tests have no items in common and the two tests are administered to different examinees, the fact that there is only one population makes the scales commensurate and the scores comparable. The usual method of assuring that the samples of examinees have been drawn from the same population is to distribute the alternative tests in rotation within a large number of classrooms at the target grade level. This method of systematic sampling can generally be depended upon to guarantee that the students responding to the two different tests belong to the same population. It requires, however, that the formats and instructions for the two tests be sufficiently similar to permit their simultaneous administration in the classroom.

3.) *Common examinee method.* This method, which is based on regression analysis, can be used only in situations where one of the tests is considered the criterion and the other the predictor. Unlike the preceding methods, it does not treat the two tests symmetrically. Both tests must be administered to the same examinee; i.e., each examinee in the sample must take both tests within a relatively short time span (typically, within a few days). The two tests are then scored by the IRT method, and regression of the criterion test scores on the predictor test scores is used to compute the criterion scores that any examinee would be expected to have obtained on the basis of his or her predictor test scores. The use of IRT scale scores for both tests will generally result in simple straight-line regression of criterion on predictor. The predicted scores will not have the same population variance as the criterion scores, but they will preserve the order relationships required to interpret the relative standings of students for purposes of cross-state comparisons.

Of these three methods, the first (common item) has the advantage of not requiring a special equating study. Data from the operational use of the tests will serve quite well. It makes the assumption, however, that the instructions and conditions of administration are identical, and that the context and placement of the common items has no effect on their probability of correct response. Cognitive test items are believed to be generally free of context effects, but empirical studies of such context effects are lacking. Identity of testing conditions might be difficult to assume when the tests are administered by independent workers in different states.

This method also assumes that the items of both tests represent the same latent dimension. Such an assumption would be justified if the item format and content of the two tests were very similar, but could be questionable if quite different approaches were taken in testing similar objectives or proficiencies. Unfortunately, there is no provision for checking on this assumption in the course of the item calibration. The main practical disadvantage of this method is that prior provision must be made for including a number of common items in all of the scales that are to be equated. Fortunately, only a few such items are needed; simulation studies (Lord, 1980) show that four to eight items will provide a dependable linkage if the samples of examinees are very large.

The second (common population) method has the disadvantage of requiring a special equating study that cannot easily be combined with operational use of the tests. But it does not require common items, and if based on a within-classroom rotation sample, it guarantees equivalence of the test instructions and conditions. This method, however, makes the same strong assumptions about a common latent dimension as the common item method.

The common examinee (regression) method is most conservative in the sense of making the fewest assumptions. Its essential requirement, namely, that one of the tests be considered a standard which the other test is attempting to reproduce, could be easily met when equating state assessment results. Since it would be impractical to attempt to equate results from all possible pairs of states, the obvious approach would be to adopt a single standard test, some or all of the scales of which the various state assessments would try to predict with one or more of their own scales. Such predictions would be required only in the direction of the standard test.

To apply this method, each of the participating states would administer the standard test to a sample of students at about the same time as the operational assessment testing. If the standard test were administered on a different day, operational testing would not be interfered with. Only a sample of some 400 to 500 students drawn from the state population at the target grade level would have to take both tests. To counter-balance order effects, half the students should take the standard test first and half the operational test first. Both tests would then be scored according to their own procedures and regression analyses performed to provide the best prediction of the standard in the least-squares sense. The states could then be compared in terms of the scales of the standard tests. These scales might not include all of those used for purposes internal to the state, but they would presumably cover those outcomes that would be relevant to national discussion of educational problems and issues.

The main advantage of the regression method in this application is that it makes no assumption about the mutual homogeneity of the two tests. Instead, it provides a multiple correlation coefficient measuring the extent to which the predicting test accounts for the variation in the criterion test scores. The two tests could be quite different in format and conditions of administration and still show a high level of correlation. Indeed, more than one scale from the predictor test might be used to predict a complex scale in the criterion test. These properties give the regression method much greater versatility than the other two methods in coping with state tests of widely differing design. For this reason, it appears to have the greatest potential for linking state assessment results, even though it lacks the desirable symmetry property of the common item and common population methods.

If a method based on predicting results of a standard test is to be implemented, the first problem to be solved is what that test should be. Agreement on the design of such a test would obviously require the effort of a national committee representating the testing programs of the participating states and augmented by curriculum experts from the various subject-matter fields. The test development work of the International Educational Achievement Association (IEA) provides a model for this type of cooperative test design. For the reasons we have suggested above, such a committee should not have undue difficulty in agreeing on specifications, at least in the main subject-matter areas. The Council of Chief State School Officers has, in fact, recognized the utility of such a broadly agreed-upon standard test, and has begun to consider steps that might be taken in its construction (CSSO, 1984, 1985).

That the instruments used by the National Assessment of Educational Progress (NAEP) could be used for this purpose is a possibility that also merits consideration.

The main impediment to their use is that, up to now, the content, types of test items, types of scales, and reporting categories used by NAEP are rather different from those typical of state testing programs. From its inception, NAEP was conceived of as an independent effort to measure educational outcomes at the national level exclusively, and it has had no significant input from the state programs in its design or implementation. If NAEP were to reform so as to be useful both at the national and state levels, it would be the obvious and natural choice for a well-recognized standard to which the state results could be related in order to provide accurate and detailed comparisons of educational outcomes in the states.

8. Conclusion

Our analysis of the potential users of data on educational outcomes,—students, parents, teachers, school counselors, school administrators, boards and officials to curriculum experts, textbook writers, state legislators and departments of education, and educational research specialists—leads us to conclude that currently existing programs for evaluating educational productivity should, and can, be redesigned to serve the needs of this varied community. We propose for this purpose the introduction of the so-called "duplex design" that supplies achievement scores for individual pupils in the main areas of subject matter, while at the same time evaluating the progress of schools in attaining the detailed objectives of the instructional program and curriculum. Based on new developments in educational statistics and measurement, including item response theory, matrix sampling, and two-stage testing, the duplex design is capable of delivering this range of information with no greater demand on testing resources and classroom time than is now required in conventional every-pupil achievement testing. As an added benefit, the scale scores in which the duplex designed assessment results are reported support both criterion-referenced and norm-referenced interpretation. They also facilitate the equating of assessment results from independent testing programs and thus provide a basis for comparison of educational outcomes across states and with reference to national and international surveys such as the National Assessment of Educational Progress and the International Association for the Evaluation of Educational Achievement.

REFERENCES

Bock, R.D. & Muraki, E. (1986). Detecting and modeling item parameter drift (in preparation).

Bock, R.D., Sykes, R.C., and Zimowski, M.F. (1986). A field trial of a two-stage assessment instrument (in preparation).

Burstein, L., Baker, E.L., Aschbacher, P. & Keesling (1985). *Using state test data for national indicators of educational quality: a feasibility study.* Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education.

CSSO. (1984). Education evaluation and assessment in the United States: Position paper and recommendations for action. Washington (D.C.): Center on Assessment and Evaluation, Council of Chief State School Officers.

CSSO. (1985). Draft report of Committee on Coordinating Educational Information & Research. Washington (D.C.): State Education Assessment Center, Council of Chief State School Officers.

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 (N.J.): Earlbaum.

Mislevy, R.J. (1985). *Inferences about Latent Populations from Complex Samples.* NAEP Research Report, 85-41. Princeton: Educational Testing Service.

Walker, D. & Schafforzich, J. (1974). Comparing curricula. *Review of Educational Research*, **44**, 83-113.

Winfield, L.F. (1986). The relationship between minimum competency testing and students' reading proficiency: Implications from NAEP (unpublished manuscript).

APPENDIX
ANALYSIS AND SCORING OF THE DUPLEX DESIGN

Virtually all item response theoretic (IRT) models in current use are defined and applied at the level of individual persons: a latent or unobservable variable characterizing a person is combined with one or more parameters characterizing a test item through a mathematical formula to give the probability that the person will answer the item correctly (Lord, 1980).

An exception is the measurement model that has been employed since 1979 by the California Assessment Program (CAP) (Bock & Mislevy, 1981). The CAP measurement model expresses the probability of correct response at the level of schools rather than at the level of individual pupils. As in the more familiar person-level IRT models, items are characterized by parameters expressing the regression of a correct response on a latent proficiency variable. The proficiency variable pertains to schools, however, and the model gives the probability of a correct response to an item with given parameters from a pupil selected at random from a given school with a given proficiency level.

Data collection under the CAP model differs radically from conventional test administration. The CAP scales, or "skill elements", are defined quite narrowly; the Grade 3 assessment of reading, for example, is comprised of sixteen separate skill elements. A pupil is administered one of twenty-five assessment booklets, containing one item each from a number of skill elements. Rather than taking a number of items from a scale to provide a basis for a score, then, an individual pupil is administered only one item from a given element. The usual IRT assumption of local independence is thereby satisfied *at the level of schools* under the CAP model. Although each booklet has the appearance of a traditional achievement test, containing a spectrum of diverse items from one or more broadly-defined content areas, all item calibration and proficiency estimation take place within the narrowly defined skill elements, at a level of schools rather than individual pupils.

When the goal of measurement is to monitor the effects of instruction, a number of important advantages accrue from this school-level model. First, the large number of narrowly-defined scales provides very detailed feedback on school curricula. The outcomes and the tradeoffs that result from shifts in emphasis in instruction or changes in allocation of resources can be tracked at the level of detail at which they can be expected to occur. Second, the data gathering design, which solicits each of a school's responses in a given subscale from a different pupil, is a member of the class of maximally efficient, item-sampling designs for estimating the school average (Lord, 1962). Third, the generic IRT advantages of item-invariant scoring and content-referenced measurement continue to hold. That is, scores are provided on a constant scale of measurement despite additions to or deletions from the item pools, and these scores are directly interpretable in terms of expected performance on any of the items in the pool. Finally, the stability of item parameters and the integrity of scales is better maintained in the narrowly-defined content elements that are suited to group-level models.

1. The objective

While providing an effective and efficient solution to the problem of assessing the effects of schools, the group-level model does not provide for the assessment of individual pupils. The resemblance of assessment booklets to traditional achievement tests, however, suggests the possibility of attaining both types of measurement from the same data. That is, schools would receive school-level measures of performance in narrowly-defined elements for the purpose of monitoring curricular effects, and pupils would receive measures of performance in the proficiencies spanning a number of elements, each of which is represented by a single item in the pupil's test booklet. The objective of this appendix, then, is to specify a model and accompanying assumptions that meet the following requirements:

1. Within a content *element*, probabilities of correct response to specified items by pupils selected at random from a given school are given by well-defined, group-level IRT models.

2. Within a *proficiency*, probabilities of correct response to specified items and specified pupils are given by a well-defined, pupil-level IRT model.

3. The pupil- and school-level models are aggregable in the sense that the expected average of pupils' proficiency scores in a given school is equal to the expected average of that school's scores in the elements. Note that the duplex model can also be defined with classrooms rather than schools as the salient grouping. For convenience, however we shall retain the school-level terminology in the sequel.

2. The data

We assume data gathered in an idealized design replicated in K test booklets (forms). The configuration of booklets for a proficiency in one content area is illustrated in Table A1. The proficiency content is comprised of J elements; each element is represented by K items; each of K booklets contains exactly one item from each element. A given pupil is administered one booklet. Random assignment of booklets to pupils within schools is assumed in the sequel.

TABLE A1
ITEM ASSIGNMENTS FOR ONE PROFICIENCY CATEGORY OF A K-FORM DUPLEX DESIGN

	Test Forms					
Content Elements	1	2	...	k	...	K
1	Item 11	Item 12		Item $1k$		Item $1K$
2	Item 21	Item 22	...	Item $2k$	...	Item $2K$
.	.	.		.		.
.	.	.		.		.
j	Item $j1$	Item $j2$	...	Item jk	...	Item jK
.	.	.		.		.
.	.	.		.		.
J	Item $J1$	Item $J2$	...	Item Jk	...	Item JK

3. The response process

Let the item represented by element j on form k be characterized by the real number, γ_{jk}, called the threshold parameter. The probability that person i_k from group h will respond to this item correctly, and receive the item score $x_{hi_kjk} = 1$ rather than $= 0$, is governed by γ_{jk} and a realization of a random variable, Z, in the following manner:

$$x_{hi_kjk} = \begin{cases} 1 \text{ if } z_{hi_kjk} > \gamma_{jk} \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

The following model is assumed for z:

$$z_{hi_kjk} = \alpha_{jk}\theta_{hi_k} + \delta_{jk}\phi_{hi_kj} + e_{hi_kjk} \qquad (2)$$

where

i) $z \sim N(0,1)$ in the population of pupils;

ii) $\theta_{hi} \sim N(\theta_h, \sigma_\theta^2)$ within group h;

iii) $\phi_{hij} \sim N(\phi_{hj}, \sigma_{\phi_j}^2)$ within group h;

iv) $Cov(\theta_{hi}, \phi_{hij}) = 0$ and $Cov(\phi_{hij}, \phi_{hij'},) = 0$ for $j \neq j'$ within group h;

v) the group means $\theta_h \sim N(0, \varsigma_\theta^2)$ in the population of groups;

vi) the group means $\phi_{hj} \sim N(0, \varsigma_{\phi_j}^2)$ in the population of groups and

vii) $Cov(\theta_h, \phi_{hj}) = 0$ and $Cov(\phi_{hj}, \phi_{hj'},) = 0$ for $j \neq j'$ in the population of groups.

From these assumptions it follows that in the unrestricted populations of persons, $\theta \sim N(0, \sigma_\theta^2 + \varsigma_\theta^2)$ and $\phi_j \sim N(0, \sigma_j^2 + \varsigma_{\phi_j}^2)$. Because of indeterminacies of units in (2) for the θ and ϕ_j scales, we may assume without loss of generality that $\sigma_\theta^2 + \varsigma_\theta^2 = 1$ and $\sigma_{\phi_j}^2 + \varsigma_{\phi_j}^2 = 1$ for all j. Hence,

viii) $\varsigma_\theta^2 = 1 - \sigma_\theta^2$

ix) $\varsigma_{\phi_j}^2 = 1 - \sigma_{\phi_j}^2$, and

x) $e_{hi_kjk} \sim N(0, 1 - \alpha_{jk}^2 - \delta_{jk}^2)$; these residual terms e are assumed to be independent over persons, groups, elements, items, and forms.

xi) The ratio $\lambda_j = \delta_{jk}/\alpha_{jk}$ is a constant over test forms k.

The resemblance of (2) to the multiple factor model for measured variables (Thurstone, 1947) is apparent. In this special case, each response process consists of contributions from a general factor θ, an uncorrelated specific factor ϕ_j, and an independent residual term. The hierarchical structure of persons within groups and the attendant normality assumptions are as outlined above. Assumption *xi* requires the "factor loadings" on the general and appropriate specific factor of items in a given element to be in a constant ratio. While the items for the element may vary in reliability, as implied by the magnitudes of α_{jk} and δ_{jk}, the *relative* impact of two factors on the *nonrandom* portion of the process is assumed constant.

4. The person-level model for the proficiency

This section derives an IRT model defined at the level of persons, expressing probabilities of correct response in the proficiency as a whole. The data observed for a given person are responses to the items on a single, randomly-selected test form comprised of one item each from the J content element. In reference to Table A1, this model may be called the "vertical" model.

Derivation of the Model. Consider the response process variable of person i_k, responding to an item on randomly-selected form k:

$$\begin{aligned} z_{hi_kjk} &= \alpha_{jk}\theta_{hi_k} + \delta_{jk}\phi_{hi_kj} + e_{hi_kjk} \\ &= \alpha_{jk}\theta_{hi_k} + e^{\circ}_{hi_kjk}, \end{aligned}$$

where

$$e^{\circ}_{hi_kjk} = \delta_{jk}\phi_{hi_kj} + e_{hi_kjk}$$

Note that $e^{\circ} \sim N(0, 1-\alpha^2_{jk})$ in the population of persons. Then

$$\begin{aligned} P(x_{hi_kjk} = 1 \mid \theta_{hi_k}) &= P(e^{\circ} > \gamma_{jk} - \alpha_{jk}\theta_{hi_k}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{\frac{-(\alpha_{jk}\theta_{hi_k} - \gamma_{jk})}{(1-\alpha^2_{jk})^{1/2}}}^{\infty} \exp(-t^2/2)dt \\ &= \Phi[a_{jk}(\theta_{hi_k} - b_{jk})] \\ &= \Phi_{jk}(\theta_{hi_k}), \end{aligned} \tag{3}$$

where

$$a_{jk} = \alpha_{jk}/(1-\alpha^2_{jk})^{1/2} \tag{4}$$

and

$$b_{jk} = \gamma_{jk}/\alpha_{jk}. \tag{5}$$

By the definition of e°, and independence assumptions given in *iv*, *vii*, and *x*, we have that $e^{\circ}_{hi_kjk}$ and $e^{\circ}_{hi_kj'k}$ are independent for $j \neq j'$. The conditional probability of response of a given response pattern is thus obtained as

$$P[(x_{hi_k1k}, \ldots, x_{hi_kJk}) \mid \theta_{hi_k}] = \prod_j [\Phi_{jk}(\theta_{hi_k})]^{x_{hi_kjk}}[1 - \Phi_{jk}(\theta_{hi_k})]^{1-x_{hi_kjk}}. \tag{6}$$

The form of (3) and the conditional independence exhibited in (6) constitute a two-parameter normal item response model (Lord, 1952), with item parameters a_{jk} and b_{jk} given by the functions of α_{jk} and γ_{jk} shown as (4). [Note that (6) would *not* follow if a test booklet contained more than one item in a given element.]

Item Calibration. Under the assumption of random assignment of test booklets (forms) to pupils, the marginal probability of a given response pattern from the unrestricted person population is

$$\int_{-\infty}^{\infty} P(\mathbf{x} \mid \theta) g(\theta) d\theta, \tag{7}$$

where $P(\mathbf{x} \mid \theta)$ is the conditional probability of response pattern $\mathbf{x}$ as given in (6) and $g(\theta)$ is the standard normal density function. After observing the response patterns $\mathbf{x}_{ik}$ of random samples of N_k persons for each test form k, the likelihood function for a and b is the product of expressions like (7) over forms and persons within forms:

$$L(\mathrm{a},\mathrm{b} \mid (\mathrm{x})) = \prod_k \prod_i \int_{-\infty}^{\infty} \prod_j [\Phi_{jk}(\theta)]^{x_{i_k jk}} [1 - \Phi_{jk}(\theta)]^{1-x_{i_k jk}} g(\theta) d\theta. \qquad (8)$$

This expression may be maximized with respect to a and b to provide maximum likelihood estimates $\hat{\mathrm{a}}$ and $\hat{\mathrm{b}}$ by means of Bock and Aitkin's (1981) EM solution, as implemented in the BILOG computer program (Mislevy & Bock, 1983).

Estimation of Person Scores. After items have been calibrated (i.e. item parameters have been estimated), it is possible to estimate proficiency scores for individual persons. Taking item parameters as known, we may obtain maximum likelihood estimates (MLE's) $\hat{\theta}_{hi_k}$ by maximizing (6) with respect to θ, given $\mathbf{x}_{hi_k}, \mathbf{a}$, and $\mathbf{b}$, or Bayes estimates $\tilde{\theta}_{hi_k}$ by evaluating the mean of the marginal probability (7) as follows:

$$\begin{aligned} \tilde{\theta}_{hi_k} &= \mathcal{E}(\theta \mid \mathbf{x}, \mathbf{a}, \mathbf{b}, g) \\ &= \int \theta(\mathbf{x} \mid \theta) g(\theta) d\theta. \end{aligned}$$

This value is readily obtained by numerical procedures outlined in Bock and Aitkin (1981) and detailed in Bock and Mislevy (1982). Indications of the precision of estimation are available for both the maximum likelihood and Bayes estimates by standard techniques. Estimates of both types, along with indication of precision, can be computed by a number of commercial computer programs, including BILOG. While both MLE's and Bayes estimates are consistent as the number of items administered increases, the Bayes estimates may be preferable in practice so that stable estimates will be obtained from the relatively small samples of items that can be anticipated in the assessment setting (e.g., perhaps 15 per major content area).

5. The group-level model for content elements

This section derives an IRT model defined at the level of groups, addressing probabilities of correct response in a single content element. The data observed from a given group are possibly several responses to each of the K items representing element j, each from a different person. (Each person contributes only one response in the element, and that to a randomly assigned item.) In reference to Table A1, such models may be referred to as the "horizontal" models.

Derivation of the Model. Consider the response process variable of person i_k from group h, responding to the randomly-assigned item from element j appearing on form k:

$$\begin{aligned} z_{hi_kjk} &= \alpha_{jk}\theta_{hi_k} + \delta_{jk}\phi_{hi_kj} + e_{hi_kjk} \\ &= \alpha_{jk}[\theta_h + (\delta_{jk}/\alpha_{jk})\phi_{hj}] \\ &\quad +[\alpha_{jk}(\theta_{hi_k} - \theta_h) + \delta_{jk}(\phi_{hi_kj} - \phi_{hj}) + e_{hi_kjk}] \\ &= \alpha_{jk}(\theta_h + \lambda_j\phi_{hj}) + e^*_{hi_kjk}, \end{aligned}$$

where

$$e^*_{hi_kjk} \sim N[0, 1 - \alpha^2_{jk}(1-\sigma^2_\theta) - \delta^2_{jk}(1-\sigma^2_{\phi_j})].$$

The probability of a correct response to item k of element j from a person *selected at random* from a group h is thus given as

$$\begin{aligned} P(x_{hjk} = 1 \mid \theta_h, \phi_{hj}) &= \frac{1}{\sqrt{2\pi}} \int^{\infty}_{\frac{-[\alpha_{jk}(\theta_h+\lambda_j\phi_{hj})-\gamma_{jk}]}{[1-\alpha^2_{jk}(1-\sigma^2_\theta)-\delta^2_{jk}(1-\sigma^2_{\phi_j})]^{1/2}}} \exp(-t^2/2)dt \\ &= \Phi[a^*_{jk}(\phi^*_{hj} - b^*_{jk})] \\ &= \Phi^*_{jk}(\phi^*_{hj}), \end{aligned} \tag{9}$$

where

$$\phi^*_{hj} = \theta_h + \lambda_j\phi_{hj} \tag{10}$$

$$a^*_{jk} = \alpha_{jk}/[1-\alpha^2_{jk}(1-\sigma^2_\theta) - \delta^2_{jk}(1-\sigma^2_{\phi_j})]^{1/2} \tag{11}$$

and

$$b^*_{jk} = \gamma_{jk}/\alpha_{jk}. \tag{12}$$

Note that b^*_{jk} is equal to b_{jk}, the item difficulty parameter in the person-level IRT model for the proficiency area as a whole.

Let N_{hjk} be the number of responses from group h to item k of element j, and let R_{hjk} be the corresponding number correct. By the definition of e^*, the independence assumptions of *iv* and *x*, and the design of the sample, we have

$$P(\mathbf{R}_{hj} \mid \mathbf{N}_{hj}, \phi^*_{hj}) = \prod_k \binom{N_{hjk}}{R_{hjk}} [\Phi^*_{jk}(\phi^*_{hj})]^{R_{hjk}}[1-\Phi^*_{jk}(\phi^*_{hj})]^{N_{hjk}-R_{hjk}}. \tag{13}$$

The form of (9) and the conditional independence exhibited in (13) constitute a two-parameter normal item response model defined at the level of groups (Mislevy, 1983), with item parameters a^*_{jk} and b^*_{jk} given by the functions of α_{jk}, γ_{jk}, σ^2_θ and $\sigma^2_{\phi_j}$ shown as (11) and (12).

Item Calibration. Suppose that vectors of numbers-correct $\mathbf{R}_{hj}$ for given numbers of attempts $\mathbf{N}_{hj}$ are observed from a sample of groups. The marginal likelihood function of $\mathbf{a}^*$and $\mathbf{b}^*$ is given by

$$L(\mathbf{a}^*,\mathbf{b}^* \mid (\mathbf{N},\, \mathbf{R})) = \prod_h \int_{-\infty}^{\infty} P(\mathbf{R}_{hj} \mid \mathbf{N}_{hj}, \phi_j^*) f(\phi_j^*) d\phi_j^*, \tag{14}$$

where f is the (normal) density of ϕ_j^*. It may be inferred from (10), *ii*, *iii*, and *iv* that $\phi_j^* \sim N(0, \sigma_\theta^2 + \lambda_j^2 \sigma_{\phi_j}^2)$. Neither $\sigma_\theta^2, \lambda_j$, nor $\sigma_{\phi_j}^2$ is known, however, so that (14) must be maximized with respect to $\mathbf{a}^*$ and $\mathbf{b}^*$ using an arbitrary variance for the normal density f, with the appropriate rescaling following as a separate step (see Section 6.1). Following the standard convention, we may calibrate under the standard scaling, under which estimates a_{jk}^{**} and b_{jk}^{**} for each item k in element j are obtained, provisional on $\phi_{hj}^{**} \sim N(0,1)$.

Estimation of School Scores. As with person-level scores, both maximum likelihood and Bayes estimates are readily obtained by a computer program such as BILOG, which accepts group-level data. The maximum likelihood estimate $\hat{\phi}_{hj}^*$ is the value that maximizes (13) with respect to ϕ_{hj}, given the data, $\mathbf{a}_j^*$ and $\mathbf{b}_j^*$; the Bayes estimate $\tilde{\phi}_{hj}^*$ is the mean of the marginal distribution after observing the data, or

$$\begin{aligned}
\tilde{\phi}_{hj}^* &= \mathcal{E}(\phi_{hj}^* \mid \mathbf{R}_{hj}, \mathbf{N}_{hj}, \mathbf{a}^*, \mathbf{b}^*) \\
&= \int_{-\infty}^{\infty} \phi_{hj}^* P(\mathbf{R}_{hj} \mid \mathbf{N}_{hj}, \mathbf{a}^*, \mathbf{b}^*, \tilde{\phi}_{hj}^*) f(\phi_{hj}^*) d\phi_{hj}^*.
\end{aligned}$$

Indications of precision—standard errors for the MLE and posterior standard deviations for the Bayes estimate–are also readily obtained by standard procedures.

6. The interface between levels

The preceding sections have derived a person-level IRT model for a proficiency and a group-level IRT model for elements, both defined with respect to the same response process model and data collection scheme. This section explicates the linkage between levels. The first consideration is the appropriate scaling of group-level item parameters; the second is the verification of the integrity of a group-level content area score.

Scaling the Group-Level Item Calibration. Calibration of items under person-level model for the proficiency as a whole (Section 4.2) provides item parameter estimates on a scale in which $\theta \sim N(0,1)$, as specified in Section 1. Calibration of items under the group-level model for content elements (Section 5.2), however, takes place on a provisional scale in which $\phi_{hj}^{**} \sim N(0,1)$, since the "natural" scale in which $\phi_{hj}^{*} \sim N(0, \sigma_\theta^2 + \lambda_j^2 \sigma_{\phi_j}^2)$ cannot be ascertained a priori. Since neither $\sigma_\theta^2, \lambda_j^2$, nor $\sigma_{\phi_j}^2$ can be known or even estimated from the data at hand, it is clear that the rescaling required to bring the group-level calibrations onto the person-level scale must be carried out by different means.

We note first that the relationship between group-level item parameters and group scores in the provisional scale and the corresponding values in the natural scale are given by

$$\phi_{hj}^{**} = \phi_h^{*}/C_j$$

$$a_{jh}^{**} = a_{jk}^{*} C_j$$

and

$$b_{jk}^{**} = b_{jk}^{*}/C_j,$$

where

$$C_j = (\sigma_\theta^2 + \lambda_j^2 \sigma_{\phi_j}^2)^{1/2}.$$

The key to rescaling is found in (5) and (12), which show that $b_{jk} = \gamma_{jk}/\alpha_{jk} = b_{jk}^{*}$. Thus

$$\hat{C}_j = K^{-1} \sum_k \hat{b}_{jk}/\hat{b}_{jk}^{**},$$

so that

$$\hat{\phi}_{hj}^{*} = \hat{\phi}_{hj}^{**} \hat{C}_j,$$

$$\hat{a}_{jk}^{*} = \hat{a}_{jk}^{**}/\hat{C}_j,$$

and

$$\hat{b}_{jk}^{*} = \hat{b}_{jk}^{**} \hat{C}_j.$$

Verification of Aggregability. The group-level score in a content area may be calculated in two ways: (1) by averaging person-level proficiency scores over the persons in that group and (2) by averaging group-level element scores for that group over elements. We now show that in the latent space, the expected value of the second, conditional on the true value of the first, is in fact equal to the first. In other words, the pairs of values will agree on the average in the population of groups.

By assumption ii, $\mathcal{E}(\theta_{hi} \mid \theta_h) = \theta_h$. Therefore, the average true score of persons in a group under the person-level model (in which the salient latent variable is θ_i) is the group mean θ_h. The same expectation holds for estimated scores of persons if they are unbiased, a condition approximated by both MLE's and Bayes estimates as the number of items increases.

Now consider the average of the group's elements scores after appropriate rescaling, as defined above in Section 6.1. If rescaling and estimation were error-free, we would have

$$\begin{aligned} J^{-1}\sum_j \phi^*_{hj} &= J^{-1}\sum_j(\theta_h + \lambda_j\phi_{hj}) \\ &= J^{-1}\sum_j \theta_h + J^{-1}\sum_h \lambda_j\phi_{hj} \\ &= \theta_h + J^{-1}\sum_j \lambda_j\phi_{hj}. \end{aligned}$$

Thus, for a randomly chosen school h,

$$\begin{aligned} \mathcal{E}\,(J^{-1}\sum_j \phi^*_{hj} \mid \theta_h) &= \mathcal{E}\,(\theta_h + J^{-1}\sum_j \lambda_j\phi_{hj} \mid \theta_h) \\ &= \theta_h + J^{-1}\sum_j \lambda_j \mathcal{E}\,(\phi_{hj} \mid \theta_h) \\ &= \theta_h, \end{aligned} \tag{15}$$

where the final step follows from assumption *vii*. Substituting approximations for parameters, we obtain

$$\mathcal{E}\,(J^{-1}\sum_j \hat{\phi}^*_{hj} \mid \theta_h) \approx \hat{\theta}_h.$$

It is *essential* for practical application of the model that the mean of the conditional distribution $p(\bar{\phi}^*_{hj} \mid \theta_h)$ be θ_h; only then can the user be assured that the two methods of aggregating up to group-level area scores will agree on the average. It is *desirable* that the variance of the same distribution be relatively small, so that agreement will be good for all groups as well as on the average.

Empirical results from the California Assessment Program (Pandey, 1984) suggest this will be the case in practice. It is the experience of CAP that the major portion of variation among items can be accounted for by major content areas as opposed to elements within areas, at both the levels of pupils and groups. That is, $\lambda_j = \delta_{jk}/\alpha_{jk} < 1$ and $\sigma^2_{\phi_j} < \sigma^2_\theta$. Assuming these inequalities and invoking the independence assumptions *iv* and *iiv*, we obtain an upper bound to the conditional variance of interest as follows:

$$\begin{aligned} Var(\bar{\phi}^*_{hj} \mid \theta_h) &= \mathcal{E}\,[\bar{\phi}^*_{hj} - \mathcal{E}\,(\bar{\phi}^*_{hj} \mid \theta_h) \mid \theta_h]^2 \\ &= \mathcal{E}\,[\theta_h + J^{-1}\sum_j \lambda_j\phi_{hj} - \theta_h \mid \theta_h]^2 \end{aligned}$$

$$
\begin{aligned}
&= \mathcal{E}(J^{-1}\sum_j \lambda_j \phi_{hj})^2 \\
&= J^{-2}\sum_j \lambda_j^2 \sigma_{\phi_j}^2 \\
&\leq J^{-1}\sigma_\theta^2.
\end{aligned}
$$

The proportion by which the variance of $\bar{\phi}^*_{hj}$ exceeds that of θ_h can be expected in practice, therefore, to fall below the reciprocal of the number of elements that comprise the content area. As a simple numerical illustration, we would expect the standard deviation of $\bar{\phi}^*_{hj}$ in the population of schools to exceed that of θ_h by less than 5 percent for as few as 10 subareas.

7. Controlling the model over time

The preceding sections derive a hierarchical IRT model for joint person- and group-level assessment at a single point in time. This section considers the dynamic extension of the procedure to multiple points in time. Issues that must be addressed are: the definition and stability of scales over time; the definition and estimation of item parameters; and the maintenance of aggregability between levels within time points.

Defining Item Parameter Drift. Under the assumptions of item response theory, item parameters have fixed values which can be applied in combination with the parameters from persons from any subpopulation and from any point in time, to yield accurate probabilities of correct response. In particular, propensities toward correct response for different items at different points in time must follow a restrictive pattern if the IRT model is to hold; namely, the differences in propensities at different time points must be explicable in terms of different distributions of the person parameters but invariant values of item parameters. Under these circumstances, the estimation of item parameters from responses at different time points would yield estimates that differed only by a simple linear transformation, aside from the modeled calibration errors associated with estimation from a finite sample. The term "item parameter drift" has come to be applied to the situation in which propensities of correct response vary over time in the population of interest in a manner than cannot be so accomodated (Mislevy, 1982). Another way of expressing the situation is to say that the scale in question is not stable over time.

Experience with the nature of item parameter drift in large-scale, ongoing testing programs has begun to accumulate in recent years. Two key conclusions are discussed below.

The first important empirical finding that supports intuition about item parameter drift is that the more narrowly scales are defined with respect to the breath of skills or content covered, the more stable and the more resistent to parameter drift they are over time. This has been confirmed in studies of data from CAP (e.g., Mislevy & Bock, 1982), where scales are defined with considerable specificity. Differential patterns of change over time from one scale to the next, however—some increasing sharply, some flat, a few declining—imply that were the items from a broad content area calibrated together to form a single scale, that scale would *not* be stable over time.

The second finding is that when drift does occur, it is confined for the most part to parameters associated with the relative levels of difficulty of items. This has been confirmed in studies of the test of Physics Achievement from the Scholastic Aptitude Tests over a ten-year period (Bock, Cook, & Pfeiffenberger, 1985). In both tests, the hypothesis of invariant item parameters under the three-parameter logistic IRT model over the time span was rejected in favor of a model that allowed thresholds to vary over time; but further relaxation of slope and asymptotic parameters in similar manners did not appreciably improve fit to the data.

These findings hold important implications for the duplex design model. It can be anticipated that item parameter drift will be negligible within a group-level model for a single content element. Patterned after the CAP design, items within a subarea will be sufficiently homogenous to insure that, with rare exceptions, the impact of societal change and curricular modifications will affect all items similarly enough to be accounted for by shifts in the population proficiency distribution. The same cannot be expected to hold for all items in a proficiency as a whole, however.

Assumptions about Change. This subsection outlines assumptions necessary to extend the duplex model to accomodate differential patterns of change over time in different elements. In line with the results of the research reviewed above, we assume that stability over time is maintained within elements (an assumption whose satisfaction can be approximated by foresightful scale construction), but not necessarily within the proficiency as a whole. Relative shifts in proficiency in different elements are instead modeled in terms of changes in item location parameters, in a manner described by Bock and Muraki (1986). In essence, the definition of the general factor θ, as implied by threshold parameters, is allowed to vary over time so as to maintain the integrity of group-level element scales, the aggregability of the group- and person-level models, and all assumptions of Section 3 except *i* and *v*.

Suppose that changes in proficiency from time t = 0 to time t = 1 maintain the covariance structure given in Section 3, but not necessarily the average levels of either the general or specific factors. That is, assumptions *ii-iv*, *vii*, *x*, and *xi*, as applied within time points remained unchanged, but

v)* $\theta_{ht} \sim N(\mu_{\theta t}, 1-\sigma_\theta^2)$ in the population of groups with $\mu_{\theta 0}=0$, and

vi)* $\phi_{hjt} \sim N(\mu_{\phi_j t}, 1-\sigma_{\phi_j}^2)$ in the population of groups, with $\mu_{\phi_j 0}=0$.

From these follow the counterpart of assumption *i*:

i)* In the unrestricted population of persons,

$$z_{hi_kjk0} \sim N(0,1)$$

and

$$z_{hi_kjk1} \sim N(\alpha_{jk}\mu_{\theta 1} + \delta_{jk}\mu_{\phi_j 1}, 1).$$

Note that no assumptions will be required concerning covariance structures across time points. Groups may therefore exhibit characteristically distinct profiles of change over time.

The Person-Level Model for a Proficiency. Let γ_{jk0} denote the threshold parameter of item k from element j at time 0, as described in Section 3. Consider the probability of a person answering this item correctly at time 1:

$$\begin{aligned} P(x_{hi_kjk}=1 \mid \theta_{hi_k}) &= P(\alpha_{jk}\theta_{hi_k} + \delta_{jk}\phi_{hi_kj} + e_{hi_kjk} > \gamma_{jk0} \mid \theta_{hi_k}) \\ &= P(\alpha_{jk}\theta_{hi_k1} + \delta_{jk}\phi_{hi_kj1} + e_{hi_kjk1} > \gamma_{jk1} \mid \theta_{hi_k}) \end{aligned}$$

where

$$\begin{aligned} \theta_{hi_k1} &= \theta_{hi_k} - \mu_{\theta 1} \sim N(0,1), \\ \phi_{hi_kj1} &= \phi_{hi_kj} - \mu_{\phi_j 1} \sim N(0,1), \end{aligned}$$

and

$$\gamma_{jk1} = \gamma_{jk0} - \alpha_{jk}\mu_{\theta 1} - \delta_{jk}\mu_{\phi_j 1}.$$

Proceeding as in Section 4.1,

$$\begin{aligned} P(x_{hi_kjk}=1 \mid \theta_{hi_k}) &= \Phi[a_{jk}(\theta_{hi_k1} - \gamma_{jk1}/\alpha_{jk})] \\ &= \Phi[a_{jk}(\theta_{hi_k} - b_{jk1})] \\ &= \Phi_{jk1}(\theta_{hi_k}) \end{aligned}$$

where a_{jk} is as in (4), the time 0 model, but

$$\begin{aligned} b_{jk1} &= (\gamma_{jk1}/\alpha_{jk}) + \mu_{\theta 1} \\ &= (\gamma_{jk0} - \alpha_{jk}\mu_{\theta 1} - \delta_{jk}\mu_{\phi_j 1})/\alpha_{jk} + \mu_{\theta 1} \\ &= (\gamma_{jk0}/\alpha_{jk}) - \lambda_j \mu_{\phi_j 1} \\ &= b_{jk0} - \Delta_j, \end{aligned} \tag{16}$$

with b_{jk0} the location parameter at time 0 as given in (5) and Δ_j a shift in location parameters *constant over all items in the elements.*

Estimation of item parameters in the person-level model over multiple time point can be carried out by the approach described by Bock and Muraki (1984). Extending the marginal maximum likelihood equation given as (8) to address data $\mathbf{x} = [(\mathbf{x})_0, (\mathbf{x})_1, \ldots, (\mathbf{x})_T]$ from time points $t = 0, \ldots, T$, we obtain

$$L(\mathbf{a}, \mathbf{b}, (\Delta) \mid \mathbf{X}) = \prod_t \prod_k \prod_i \int_{-\infty}^{\infty} \prod_j [\Phi_{jkt}(\theta)]^{x_{i_k jkt}} [1 - \Phi_{jkt}(\theta)]^{1 - x_{i_k jkt}} g_k(\theta) d\theta,$$

where $(\Delta) = (\Delta_{jt})$ with $j = 1, \ldots, J$ and $t = 0, \ldots, T$ but $\Delta_{j0} \equiv 0$,

$$\Phi_{jkt}(\theta) = \Phi[a_{jk}(\theta - b_{jk} + \Delta_{jt})] \tag{17}$$

and

$$g_k(\theta) = (2\pi)^{-1/2} \exp[-(\theta - \mu_{\theta t})^2/2], \text{ with } \mu_{\theta 0} \equiv 0. \tag{18}$$

An indeterminacy of origin with respect to $\mu_{\theta t}$ and Δ_t is apparent in (17) and (18). Without further loss of generality, we may resolve this indeterminacy by requiring that at each time point t,

vii)*

$$\sum_j \Delta_{jt} \equiv \sum_j \lambda_j \mu_{\phi_j t} = 0.$$

That is, a change in this weighted average of specific factors is identically equal to a change in the general factor under the model specified in (2). We note in passing that the weight assigned each specific factor, $\lambda_j \equiv \delta_{jk}/\alpha_{jk}$, is proportional to the influence of the specific factor relative to the general factor upon items in that element.

The Group-Level Model for Subareas. Consider as in Section 5 the probability of a correct response to item k of element j from a person selected at random from group h, now at time t = 1:

$$P(x_{hi_kjk} = 1 \mid \theta_h, \phi_{h_j}) = P[\alpha_{jk}(\theta_{h1} + \lambda_j \phi_{hj1}) + e^*_{hi_kjk} > \gamma_{jk1}]$$

where

$$\begin{aligned}
\theta_{h1} &= \theta_h - \mu_{\theta 1}, \\
\phi_{h_j 1} &= \phi_{h_j} - \mu_{\phi_j 1}, \\
\gamma_{jk1} &= \gamma_{jk0} - \alpha_{jk}\mu_{\theta 1} - \delta_{jk}\mu_{\phi_j 1},
\end{aligned}$$

and

$$e^*_{hi_kjk} = \alpha_{jk}(\theta_{hi_k} - \theta_h) + \delta_{jk}(\phi_{hi_kj} - \phi_{h_j}) + e_{hi_kjk} \sim N[0, 1 - \alpha^2_{jk}(1-\sigma^2_\theta) - \delta^2_{jk}(1 - \sigma^2_{\phi_j})].$$

Proceeding as in Section 5.1,

$$\begin{aligned}
P(x_{hi_kjk} = 1 \mid \theta_h, \phi_{h_j}) &= \Phi[a^*_{jk}(\phi^*_{hj1} - \gamma_{jk1}/\alpha_{jk})] \\
&= \Phi[a^*_{jk}(\phi^*_{h_j} - b^*_{jk1})],
\end{aligned}$$

where a^*_{jk} is the same group-level slope parameter defined in (11) but

$$\begin{aligned}
\phi^*_{hj1} &= (\theta_h - \mu_{\theta 1}) + \lambda_j(\phi_{h_j} - \mu_{\phi_j 1}) \\
\phi^*_{hj} &= \theta_h + \lambda_j(\phi_{hj} - \mu_{\phi_j 1})
\end{aligned}$$

and

$$\begin{aligned}
b^*_{jk1} &= (\gamma_{jk1}/\alpha_{jk}) - \mu_{\theta 1} \\
&= (\gamma_{jk0} - \sigma_{jk}\mu_{\phi_j 1})/\alpha_{jk} \\
&= b_{jk0} - \lambda_j \mu_{\phi_j 1} \\
&= b_{jk1}. \qquad (19)
\end{aligned}$$

A comparison of (19) and with (16) reveals that the equivalence of person- and group-level item difficulty parameters is maintained over time under the restrictions given in Section 7.2.

Furthermore, aggregability in the sense of Section 6.2 is maintained as well:

$$\mathcal{E}(\bar{\phi}^*_{hj} \mid \theta_h) \;=\; \mathcal{E}\{J^{-1}\sum_j [\theta_h + \lambda_j(\phi_{h_j} - \mu_{\phi_j 1})] \mid \theta_h\}$$

$$\begin{aligned}
&= \theta_h + J^{-1}\mathcal{E}\big(\sum_j \lambda_j \phi_{h_j}\big) - J^{-1}\mathcal{E}\big(\sum_j \lambda_j \mu_{\phi_j 1}\big) \\
&= \theta_h + J^{-1}\sum_j \lambda_j \mathcal{E}\,(\phi_{h_j}) - 0 \\
&= \theta_h + J^{-1}\sum_j \lambda_j \mu_{\phi_j 1} \\
&= \theta_h.
\end{aligned}$$

Comments. Two properties of the dynamic extension of the duplex model merit special mention at this point.

First, the integrity of group-level element scales (and, by implication, person-level element scales, as discussed in Mislevy, 1982) is maintained in an ideal manner. Under the restrictive assumptions about the nature of change over time in each element, item parameters remain invariant and differences in the propensities of correct response to different items are explicable in terms of shifting distributions of proficiency alone.

Second, although the item *slope* parameters remain invariant in the person-level model for the content area as a whole, the item *location* parameters do not. They exhibit shifts that reflect differential patterns of change in different content elements. If performance has improved in one element, for example, but declined in a second, then the location parameters of items in the first element will be relatively lower (easier) compared to those of the second when thresholds are updated.

While time-dependent item parameters represent a distinct departure from typical practice, a choice must be made when it is desired to use IRT to model performance over a content area of sufficient breadth to invite item parameter instability. Use of a single unidimensional model without regard for the consequences of its lack of fit across time is an exceedingly poor choice. More sensible choices are (1) to model performance only within more narrowly defined scales and take averages of performance over scales, or (2) to use a single model that allows for differential trends in different item subsets implicitly, through structured changes in item parameters. The approach taken here for the duplex model combines features of both of these latter approaches.

8. Some final comments

In a review of IRT methodology for educational assessment, Bock, Mislevy, and Woodson (1982) outline two approaches well-suited to the population focus of educational assessment. The first approach is the use of the more familiar person-level models, though bypassing the computation of person-level results by estimating item and population parameters directly from counts of response patterns. The second is the use of group-level models in narrowly-defined content areas. Both approaches have advantages and disadvantages.

The first approach, based on person-level models, shares two key features with the duplex design: (1) scales are defined narrowly in order to enhance their stability over times, and (2) data are collected in efficient designs that proscribe estimation for individuals at the level of the elemental scales. Application of such an approach requires marginal estimation procedures for item parameters (e.g., Bock & Aitkin, 1981), and for population characteristics (e.g., Mislevy, 1984). An integration with survey research methodology for finite populations and complex sampling designs is also available (Mislevy, 1985). This approach imposes fewer assumptions than the second, but is more burdensome computationally and, by requiring multiple responses from a respondent in a scale, provides less efficient estimates of group-level performance when the assumptions of the group-level model are met.

The duplex model presented in this paper is a logical extension of the second approach if information about individuals in a more broadly-defined content area is also desired. It maintains the group-level model's advantage of maximum efficiency for group-level results, and imposes less a computational burden than the first approach. This is achieved at the cost of more restrictive distributional assumptions, such as homoscedasticity within groups and over time. There is also less connection to traditional methodologies of survey sample research. Rather than estimating characteristics of the finite populations that groups constitute, the model presented here explains performance as a manifestation of processes under the control of a latent structure, and estimates the parameters that characterize the structure. As such, it shows a greater affinity for "superpopulation" models in survey research (e.g., Royall, 1970).

References

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

Bock, R.D., Cook, L., & Pfeiffenberger, W. (1985). *Detecting and Modeling Item Parameter Drift.* Paper presented at the 50th anniversary meeting of the Psychometric Society, Nashville, TN.

Bock, R.D., & Mislevy, R.J. (1981). An item response curve model for matrix-sampling data: The california grade 3 assessment. In D. Carlson (Ed.), *Testing in the States: Beyond Accountability.* San Francisco: Jossey-Bass.

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

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

Bock, R.D., & Muraki, E. (1986). Detection and modeling of item-parameter drift (in preparation).

Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph No. 7*. Psychometric Society.

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: Lawrence Erlbaum Associates.

Mislevy, R.J. (1982). *Toward an Understanding of Item Parameter Drift.* Paper presented at the annual meeting of the American Eduational Research Association, New York.

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

Mislevy, R.J. (1984). Estimating latent populations. *Psychometrika*, **49**, 359-381.

Mislevy, R.J. (1985). *Inferences about Latent Populations from Complex Samples.* NAEP Research Report, 85-41, Princeton, NJ: Educational Testing Service.

Mislevy, R.J., & Bock, R.D. (1982). *Stability of Item Parameters in the CAP Grade 3 Assessment.* Sacramento (CA): Division of Planning, Evaluation & Research, State Department of Education.

Mislevy, R.J., & Bock, R.D. (1983). *BILOG: Item Analysis and Test Scoring with Binary Logistic Models* Mooresville, Ind: Scientific Software.

Pandey, T. (1984). Personal Communication.

Royall, R.M. (1970). On Finite Population Sampling Theory under Certain Linear Regression Models. *Biometrika*, **57**, 377-387.

Thurstone, L.L. (1947). *Multiple Factor Analysis.* Chicago: University of Chicago Press.