INDIVIDUALIZED EDUCATIONAL ASSESSMENT: TWELFTH-GRADE SCIENCE

CSE Technical Report 324

R. Darrell Bock & Michele Zimowski

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

June 1991

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

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

As originally conceived in the National Assessment of Educational Progress (NAEP), educational assessment was intended to report educational outcomes at a high level of aggregation—average attainment for states, regions, or the nation. Using efficient techniques of multiple matrix sampling in which each student responds to only a limited number of items randomly selected from a much larger set, NAEP attained high levels of generalizability for numerous educational objectives with relatively small demand on student time. When this technique was adapted to state-level assessments, as in the California Assessment Program, the reporting was extended to the level of separate schools, but there was no attempt to evaluate the attainment of individual students.

Although the state programs based on this conception served well the needs of policymakers, planners, and curriculum specialists, they did not satisfy the requirements of school principals, teachers, and parents for information that would guide and certify the progress of individual students. Neither did they motivate students by giving them a personal stake in the outcome of their efforts on the assessment tests. To satisfy these additional demands on the testing program, many states adopted, at the expense of duplication of effort and further encroachment on classroom time, a two-tiered program including matrix-sampled assessment testing and traditional student-level achievement testing. In an effort to serve the purposes of these overlapping testing program in a single, comprehensive assessment, the National Opinion Research Center (NORC) in 1986 began studies of a new approach to educational evaluation that combined features and benefits of school-level matrix sampling assessment and individual-level achievement testing.

The NORC Assessment Design Project

With support from the U.S. Office of Educational Research and Improvement, NORC developed and field-tested in the states of Illinois and California a new type of multipurpose assessment instrument based on the "Duplex Design" of Bock and Mislevy (1988). The first trial of the design applied to eighth-grade mathematics and provided student-level scores in the content areas of Number, Algebra, Geometry, Measurement, and Statistics, and in the process skill areas of Factual Knowledge, Conceptual Understanding, and Problem-solving; at the same time it provided school-level scores in 45 curricular objectives in eighth-grade mathematics. The results of the field trials clearly demonstrated that the duplex instrument could provide detailed program evaluation as well as accurate scores for students, schools, districts and the state, all based on a testing session intermediate in length between matrix-sampled assessment and traditional achievement testing (Bock & Zimowski, 1989).

Since August 1989, at the invitation of the National Science Foundation and with continuing support from OERI, NORC has been implementing another duplex design in the areas of Earth Science, Biology, Chemistry, and Physics at the twelfth-grade level. Exploiting new technology for computer-controlled laser printing and optical character reading, this project has achieved a breakthrough in large-scale assessment technique by producing test forms individualized to the course background and performance levels of each student taking the science test. These forms include multiple-choice and open-ended essay questions. In addition, the assessment design incorporates a similarly individualized component of "hands-on" laboratory performance assessment in General Science, Biology, Chemistry, and Physics. The project has also led to the development of materials and procedures for a "Graded Mark-point" method of reliably scoring extended responses the open-ended and laboratory performance exercises.

The present report describes the goals, principles, and methods of the individualized educational assessment as implemented in a twelfth-grade science

assessment instrument now undergoing field trials in the state of Ohio. A pilot school in Ohio will be tested in the first week of December, 1990, and all twelfth-grade students in a stratified probability sample of 40 Ohio schools will be tested in March and April of 1991. A final report of the study is due on November 30, 1991.

Two-stage Testing

In addition to the duplex principle, the other major innovation of the NORC assessment design project is the practical implementation of two-stage testing. Since early in the development of item response theory (IRT), it has been known that substantial reductions in testing time, without sacrifice of test reliability, can be obtained by a form of adaptive testing in which students are tested in two stages, where the second-stage test is selected to be maximally informative, given the student's score on the first-stage test. Studies by Lord (1980) showed that with second-stage forms representing at least four levels of difficulty, comparable reliability could be obtained in about one-third the testing time required for a conventional achievement test.

For a long time the logistics of two-stage testing were thought to be too complex to allow applications in large-scale assessment programs. Recently, however, two technological developments have radically changed this picture. One of these is the availability of high-capacity, programmable optical character readers used commercially in processing responses from direct-mail advertising promotions. The readers are capable of scoring test booklets duplicated by any printing method, rather than the high-precision printing previously required for scannable test booklets. The other development is that high-capacity laser printers driven by computers are now able to assemble the material to appear in a test booklet as the pages are printed. The NORC assessment design project has made use of this technology to implement large-scale two-stage testing in a practical way.

First-stage Test

The first-stage test booklet is designed to be administered in February. It consists of a student questionnaire, asking for high school course history in Science and Mathematics, and a 20-item pretest with 5 items each in the areas of Earth Science, Biology, Chemistry and Physics. The items of the pretest are widely spaced in difficulty and give a rough estimate of the student's level of proficiency in these subjects. On the basis of a student's response to course background questionnaires and the score on the pretest, he or she is assigned a second-stage form adapted to an appropriate level of science preparation in each of the four areas. The questionnaire and test are designed to be administered to twelfth-grade students by teachers in the participating schools. The completed test booklets are returned to NORC for scanning and analysis, and the results are used to control the generation of the second-stage forms appropriate for each student. Each such booklet is labeled clearly with the student's name on the cover, and each page of each booklet also carries optically readable numbers that identify the student and the items of that particular form.

Second-stage Test

The second-stage test is designed to be administered in late March or early April of the student's twelfth-grade program. The forms of the test are of two types, which can be administered separately or in combination: Type I consists of only multiple-choice items; Type II consists of multiple-choice items and open-ended items.

The Type I forms are further divided into a part A and part B, each consisting of 32 items. If the test is to be used to assign scores to students for purposes of certification, it is recommended that each student be administered both part A and part B in 80 minutes of testing time. If the scores are to be used only to evaluate schools or programs, or to inform interested parties, each student may be randomly assigned a part A or part B, to be administered in 40 minutes of testing time.

The Type II forms each contain 32 multiple-choice items and 4 open-ended items. Forty minutes of testing time is allocated to the multiple-choice items in the first half of the form, and 40 minutes for open-ended items in the second half of the form. These forms are also divided into a part A and part B consisting of 16 multiple-choice items and 2 open-ended items. These parts may also be administered separately if highly accurate student-level scores are not required.

Forms and Booklets

Each second-stage test form, including part A and part B, is replicated in parallel six times. In addition, each form consists of four booklets constructed at each of four levels of difficulty: the lowest level is aimed at students who have only one course in secondary-school science; the next two levels are aimed at students with at least two courses in science, with the lower level being assigned to those students who score below the median on the pretest and the higher level being assigned to students who score above the median; the highest level of difficulty is aimed at students with Advanced Placement courses of science.

Because the second-stage forms are produced by computer contingent on information from the student questionnaires and pretest, the relative difficulty can be adapted to the type of course background of each student. For example, if a student has one course in earth sciences, two in biology, and one in chemistry, the biology content might be pitched at level three, the chemistry at level one, and the physics and earth sciences at level two. If a student has only one course in general science or earth sciences and biology, but has a reasonably good pretest score, the second-stage test will be pitched at level two in biology and earth sciences, but at level one in chemistry and physics. All possible profiles of student preparation can be accommodated by these computer-generated second-stage forms.

Item Structure of the Second-stage Forms

The content-by-process classification of items in the second-stage form is shown in Table 1. The table represents the items of the 64-item form. The open and cross-hatched entries represent one of the possible divisions of the form into part A and part B. Other forms select item- and process-content for part A and part B in all possible combinations. Each test form is a random assignment of items classified according to the categories of Table 1. From a pool of 11,500 items, 24 test booklets have been constructed (6 forms at each of 4 levels of difficulty). Thus, the second-stage instrument consists of stratified randomly parallel forms containing a possible 1,536 different items.

IRT Scaling

The instrument is based on a Duplex Design intended for scoring on IRT scales in three directions. At the student level, scale scores can be computed for (a) each of the four content areas and (b) each of the four process categories, for a total of eight scales, plus an overall index of science achievement. At the school level,

TABLE 1.
Content-by-Process Item Classification

Content	Part A ☐ Part B ☒			
1. Physics	*****	*****	*****	*****
Mechanics	X			X
Electricity and Magnetism	X		X	
Heat and Kinetic Theory		X		X
Waves, Optics, and Sound		X	X	
2. Chemistry	*****	*****	*****	*****
The Atomic Model	X			X
Chemical Reactions	X		X	
Quantitative Chemistry		X		X
States of Matter		X	X	
3. Biology	*****	*****	*****	*****
of the Cell	X			X
of the Organism	X		X	
Reproduction and Genetics		X		X
Biological Diversity		X	X	
4. Earth Sciences	*****	*****	*****	*****
Space	X			X
Air	X		X	
Water		X		X
Land		X	X	
Process	Knowledge of Scientific Terminology and Facts	Knowledge of Scientific Methods and Procedures	Understanding of Scientific Concepts and Principles	Problem Solving