ED482268 2003-09-00 Matrix Sampling of Test Items. ERIC Digest.

ERIC Development Team

www.eric.ed.gov

Table of Contents

If you're viewing this document online, you can click any of the topics below to link directly to that section.

ERIC Identifier: ED482268
Publication Date: 2003-09-00
Author: Childs, Ruth A. - Jaciw, Andrew P.
Source: ERIC Clearinghouse on Assessment and Evaluation College Park MD.

Matrix Sampling of Test Items. ERIC Digest.

THIS DIGEST WAS CREATED BY ERIC, THE EDUCATIONAL RESOURCES INFORMATION CENTER. FOR MORE INFORMATION ABOUT ERIC, CONTACT ACCESS ERIC 1-800-LET-ERIC

"THE CHALLENGES OF TEST DEVELOPMENT"

Imagine that you must create a test of science knowledge and skills to be administered to all fifth-grade students in your state or province. Based on the test results, reports of individual students' mastery of the curriculum will be sent to parents and teachers. Summary reports will also be sent to schools and school districts to help them evaluate how well they are teaching the curriculum. You and your staff review relevant curriculum

documents and compile a list of the things fifth-grade students should know and be able to do. Your team begins to develop test items about the parts of the human circulatory system, what happens when water freezes, why a pulley system works, how clouds form, and so on. Most of the items you develop require the students to construct and justify their responses. Only a few items are multiple-choice.

After developing and pilot testing a large number of items, you begin to assemble the test. The pilot test showed that each constructed-response item takes about 10 minutes to complete. The multiple-choice items take an average of 2 minutes. You and your staff create a test that samples all areas of the science curriculum. It has 32 constructed-response items and 16 multiple-choice items. If your time estimates are correct, the test will require almost 6 hours, plus time for the instructions, warm-up, and breaks. The fifth-grade students will also be taking tests in other subject areas, so the total testing time will be several times that.

You must decide what to do. You are being pressured to reduce the testing time to 2 hours, including instruction time and breaks. Your item writers, however, argue that a test with fewer items will not adequately cover the curriculum. With fewer items, whole sections of the curriculum might be omitted. Teachers and students might conclude that, because they are not on the test, those parts of the curriculum are less important.

You consider replacing some of the constructed-response items with more multiple-choice items. A mostly multiple-choice test could cover more content in less time. However, you worry that multiple-choice items may fail to test the students' depth of understanding and skill in applying knowledge. Such a test might cover more of the curriculum, but superficially.

Each of the alternatives you consider requires a compromise. Adequate content coverage, but too much testing time. Less testing time, but inadequate content coverage. Faster items, but a lower quality assessment. You reason that your testing program cannot be the only one facing these choices. What are other programs doing? Are there other alternatives?

"THE CONCEPT OF MATRIX SAMPLING"

One approach to achieving broad curriculum coverage while minimizing testing time per student is matrix sampling of items. Matrix sampling involves developing a complete set of items judged to cover the curriculum, then dividing the items into subsets and administering each student one of the subsets of the items. Matrix sampling, by limiting the number of items administered to each student, limits the amount of testing time required, while still providing, across students, coverage of a broad range of content.
A word about terminology: Popham (1993) labels the type of matrix sampling just described item sampling. It is also possible to sample students, so that only some of the students at a grade level take any test at all. This approach is used for the National

Center for Education Statistics' National Assessment of Educational Progress in the United States. And, of course, both items and students can be sampled an approach that Popham calls genuine matrix sampling. Sampling of students may be possible in some testing programs, but many require testing of all students. The recently enacted No Child Left Behind legislation, for example, requires that all U.S. students in grades three through eight be tested annually in reading and mathematics.

For the science test just described, the 32 constructed-response items and 16 multiple-choice items could be divided into four sets of items, each with eight constructed-response items and four multiple-choice items. Each student could be randomly assigned to take only one of the four sets of items. In this way, testing time could be held to less than two hours and, across the four sets of items, the curriculum would be adequately covered. Of course, the compromise would be that comparing results across students would require extra work and might be difficult to explain to the public. However, aggregated results at the school, district, and state/provincial levels would be based on the full set of items that covered the curriculum.

A variation of matrix sampling helps with the problem of comparing results across students. This variation is sometimes called partial matrix sampling. After a set of items has been developed to provide adequate coverage of a content framework, a subset of those items is selected to form the "common" items administered to all the students. The remaining items are then matrix-sampled. Each student receives a form that combines the common items with some matrix-sampled items. The common items help to improve the comparability of student results, while the matrix-sampled items increase content coverage per testing time (Dings, Childs, & Kingston, 2002). For the science test, for example, four common constructed-response items could be chosen and the remaining 28 constructed-response items divided into seven sets of four items each. Similarly, the multiple-choice items might be divided into two common items and seven sets of two items each.

"COSTS OF MATRIX SAMPLING"

Two issues that must be considered when deciding what design to use in a testing program are content coverage and testing time. Additional considerations include such issues as printing and scoring costs and the precision of student- and group-level scores. These considerations can be thought of as different types of costs. The companion Digest, Costs of Matrix Sampling of Test Items, presents nine categories of costs more fully: development costs, materials costs, administration costs, educational costs, scoring costs, reliability costs, comparability costs, validity costs, and reporting costs.

With unlimited resources, all costs could be met and an optimal plan could be implemented. However, resources are not unlimited. Every test design we consider, therefore, involves a compromise. The various types of costs must be considered jointly for two reasons. First, the costs are different in both kind and extent, but are

interrelated. Limiting spending in one area may lead to costs in another area. For example, developing fewer items may reduce development costs, but also reduce validity a cost that should not be ignored. Second, the costs may not be equally important. Some expenses may be more tolerable than others. For example, if the stakes of a test are very high, then the reliability of the test will be very important and other costs may be determined relative to a target reliability. If we need to derive both student- and school-level scores, then that must be considered in selecting a test design. The categories of costs should be considered with their inter-relatedness and relative importance in mind.

"EXPLORING THE VIABILITY OF MATRIX SAMPLING OF ITEMS"

How should state or provincial testing officials proceed if they are considering using matrix sampling of items? As outlined in the previous section, every test design, whether or not it involves a matrixed component, carries with it certain costs. The various costs will be of differing levels of importance for different testing programs depending on their circumstances. A testing program would want to examine the costs in light of its mandate(s), the content of the tests, and the financial resources available, among other considerations when choosing a design.
Clearly, a state's or province's choice of test design requires careful consideration of the various costs associated with each possible design in relation to the testing program's goals and constraints. Ideally, estimates of the reliability, comparability, and validity costs could be based on pilot studies within the state or province or on data from similar jurisdictions. Because every design represents a compromise in terms of one or more costs, only by considering the various costs together can we hope to make the best decisions.

REFERENCES

Dings, J., Childs, R., & Kingston, N. (2002). The Effects of Matrix Sampling on Student Score Comparability in Constructed-Response and Multiple-Choice Assessments. Washington, DC: Council of Chief State School Officers.
Popham, W. J. (1993). Circumventing the high costs of authentic assessment. Phi Delta Kappan, 7, 470-473.

"ADDITIONAL READING"

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. Educational Measurement: Issues & Practices, 14, 9-12, 27.

Fitzpatrick, A. R., Lee, G., & Gao, F. (2001). Assessing the comparability of school scores across test forms that are not parallel. Applied Measurement in Education, 14, 285-306.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. Applied Measurement in Education, 7, 323-342.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In Technical Issues in Large-Scale Performance Assessment. Report No. NCES 96-802 (pp. 59-78). Washington, DC: U.S. Department of Education.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. Journal of Educational Measurement, 29, 133-161.

This publication was prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education, under contract no. ED-99-CO-0032. The opinions expressed in this report do not necessarily reflect the positions or policies of OERI, or the U.S. Department of Education

—

Title: Matrix Sampling of Test Items. ERIC Digest.
Document Type: Information Analyses---ERIC Information Analysis Products (IAPs) (071); Information Analyses---ERIC Digests (Selected) in Full Text (073);
Available From: ERIC Clearinghouse on Assessment and Evaluation, 1120 Shriver Laboratory, University of Maryland, College Park, MD 20742. Tel: 800-464-3742 (Toll Free). Web site: http://ericae.net.
Descriptors: Item Banks, Matrices, Sampling, Test Construction, Test Content, Test Items, Timed Tests
Identifiers: ERIC Digests