ED482270 2003-09-00 Costs of 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: ED482270
Publication Date: 2003-09-00
Author: Childs, Ruth A. - Jaciw, Andrew P.
Source: ERIC Clearinghouse on Assessment and Evaluation College Park MD.

Costs of Matrix Sampling of Test Items. ERIC Digest.

Matrix sampling of items that is, division of a set of items into different versions of a test form is used by several large-scale testing programs. Like other test designs, matrixed designs have advantages and disadvantages. For example, testing time per student is less than if each student received all the items, but the comparability of student scores may decrease. Also, curriculum coverage is maintained, but reporting of scores becomes more complex. In this Digest, nine categories of costs associated with matrix sampling are discussed: development costs, materials costs, administration costs, educational costs, scoring costs, reliability costs, comparability costs, validity costs, and reporting costs.

Development costs include the cost of writing items, subjecting them to sensitivity and technical reviews, pilot and field testing them, and analyzing the pilot and field test results. In general, developing more items requires more staff time and more participation by schools. In small jurisdictions, developing large numbers of items may be particularly burdensome because the cost of developing additional items raises the per-student cost of testing more quickly when there are fewer students taking the test; and the numbers of schools and students available to pilot and field test new items are limited.

Materials costs include the expense of printing the test booklets and of shipping them to schools. Longer tests are more expensive to print and, because the resulting booklets are larger and heavier, shipping costs more. In addition, although new computerized printing technologies are helping to decrease the costs of printing multiple versions of the test booklets, the complexity of preparing multiple versions for printing still means that they are more expensive to produce than a single version of a test.

If a test is to be administered by computer, the materials costs are very different, of course. Instead of printing and shipping test booklets, test developers must procure or arrange for the use of computers and must set up a computer program to deliver the test. Depending on the computerized administration approach, the costs of administering different versions of the test may or may not be greater than administering a single version.

Another possibility is to print a single version of a test booklet and instruct different students to respond to different sections of the booklet. This avoids the costs of printing multiple versions of the test booklet, but results in a longer booklet, and perhaps, some confusion for students.

In addition to test booklets, other materials must be printed and mailed to the schools, including instructions for handling the test booklets and administering the test and explanations of why the test is being given and how the results of the test should be interpreted. Parents may receive materials, either directly or through the schools, explaining the purpose of the test and reporting their child's results. Some of these materials may be distributed via the Internet, but printed materials are still required for parents without Internet access.

Administration costs include the time teachers and other school personnel must devote to preparing for the test administration (reviewing the procedures and sorting the materials), administering the test, and returning the materials to the scoring site. Depending on the complexity of the test administration, the time required can vary widely. If the test is short, requires only paper and pencil, and has only a single version, the time to prepare for the administration may be minimal. However, if the test is longer, if students must perform tasks in addition to writing responses, and/or if different test versions must be administered in a complex pattern, then the time required to prepare

for the test administration can be quite substantial. Multiple test versions, by themselves, need not increase administration costs, however, if the booklet distribution pattern is very simple.

Educational costs include the time that is taken from other educational activities for test preparation and administration. They also include the impact of knowing that the test will be administered on the way teachers cover the curriculum. For example, teachers may increase the amount of time they spend teaching parts of the curriculum they expect to be on the test and decrease the amount of time spent on other concepts. The test may also impact the way that a district or school allocates resources. For example, resources may be directed disproportionately to the grade levels that will have to take the test, if the district or school believes that doing so might improve test results.

Scoring costs include the costs of scanning and processing "bubble sheets" for multiple-choice items, preparing test scoring guides, recruiting and training judges to mark student responses to constructed-response items, completing the scoring, and checking and processing the results. Scanning multiple-choice responses is relatively inexpensive. However, recruiting and training judges to mark students' test responses is time-consuming and expensive. As the amount of student work to be marked increases, the amount of scoring, of course, increases. Even if the length of the test remains constant, if there are multiple versions of the test, then the costs of preparing scoring guides, of training judges, and of processing results will increase because the number of unique items across test versions will increase.

The financial and logistical costs of scoring more student work may be partly offset, however, if one purpose of the testing program is to provide scoring-related professional experience and employment for teachers and/or others. If the amount of scoring increases, the amount of employment available will increase.

Reliability costs. Reliability refers to how accurate and how consistent scores are. Some test designs lead to more accurate and consistent scores than other designs. The type of accuracy and consistency of interest will depend on the type of score that the test must yield. Is the test intended to produce a score on a scale of 1 to 100? If so, it may be important to know the confidence interval or standard error around the student's score. Is the test's purpose to place students in one of several performance levels or yield a pass-fail decision? In either case, the accuracy of the decision would be important.

Different levels of scores must also be considered. Many testing programs are required by law to produce both student scores and school- or district-level scores. In addition, they may provide summary results for the entire jurisdiction. Paradoxically, some test designs can increase the reliability of one score level while decreasing reliability at another level. In particular, if the number of items an individual student answers is small, the reliability of student scores will be low. However, if multiple forms of the test are

administered, the number of items contributing to the school- or district-level score may be large. As Shoemaker (1971) explains, a classical test theory analysis of the resulting data would yield a mean test score for each group of students who happened to take the same items, and the mean school test score would be computed as a weighted composite of the subgroup scores. The standard error of the school mean test score based on the matrixed test would be smaller than the standard error from a test of the same length, but in which all student scores were based on the same items. In an item response theory (IRT) analysis of the same data, however, administering different items to different students would not necessarily improve score reliability for students, schools, and districts. To do so without increasing the number of items per student, the test would have to become "adaptive" that is, it would avoid administering easy items to those students who have better mastery of the material and are almost certain to get those items right, and avoid administering hard items to students who are almost certain to get them wrong.

Comparability costs. It is usually assumed that the scores of different students taking a test can be compared. Comparability is improved by uniform administration conditions and equivalent marking. It also can depend on the particular items that students receive. If all students receive the same items, then their scores are easier to compare than if they receive different items. The comparability of aggregate scores, such as school- or district-level results, is also important to consider.

The approach chosen to analyze the test results makes a difference. If the items are calibrated or equated onto a single scale using item response theory, whether a student answered the same or different items should have little effect on comparability. However, if classical test theory is used, then the particular items may affect comparability. IRT models require at least several hundred responses per item. The comparability of performance assessment results, whether reported at the student level or in aggregated form, has been addressed by a number of authors (e.g., Bock & Mislevy, 1987; Brennan & Johnson, 1995; Cronbach, et al., 1995; Fitzpatrick, Lee, & Gao, 2001; Haertel & Linn, 1996; Mislevy, et al., 1992). Haertel and Linn (1996), for example, write:

"Consider the case of a state-level testing program that administers different sets of items to different students in order to improve school-level achievement estimates, but which also produces individual-level scores. Unless students' scores are based solely on items administered to all of them in common, some degree of comparability must be assumed across the items given to different students. (There is a dilemma here. The more comparable the matrix-sampled items are, the less matrix sampling improves content coverage.) (p. 64)."

In other words, the types of items that will most improve the meaningfulness of school-level results may well decrease the comparability of student-level results.

Computing statistics to measure the comparability of students' scores can be quite complex. Cronbach, et al. (1995; see also Brennan & Johnson, 1995, for a similar discussion) propose an approach to examining the standard errors of results at the student- and school-level, using generalizability analysis. As they point out, the comparison of scores for individual students who do not take the same test form requires the computation, not just of the standard error of the scores, but the standard error of the difference between the scores, which is likely to be considerably larger.

Validity costs. Validity refers to the extent to which a test is measuring what it is intended to measure. Different test designs may be more or less valid for different uses and interpretations. A particular concern for the validity of a test intended to measure mastery of a school curriculum is how well the test represents the curriculum. If the test includes a large number of items sampled from across the curriculum, then there is a better chance that the test reflects mastery of the curriculum than if the test includes a very few items and so omits large sections of the curriculum.

The degree to which a test measures the intended construct can also be affected by how easily students are able to demonstrate their knowledge on the test. For example, confusing test instructions may interfere with students' ability to demonstrate their knowledge. Fatigue may also impact student scores if the test is very long, interfering with how well the test measures student mastery. Other sources of bias include a test's reading level or the inclusion of extraneous concepts that may be less familiar to some students than to others.

Reporting costs. A more complex test design may require more explanatory materials and more communication with educators, parents, and the media. This is especially true if the complex design supports certain scores at some score levels (e.g., at the school- or district-level) and not at others.

The nine categories of costs will vary in importance depending on the testing program. Testing directors and their staffs must examine relevant costs in light of their mandate(s), the content of the tests, the financial resources available, and the acceptability of the inevitable compromises.

REFERENCES

Bock, R. D., & Mislevy, R. J. (1987). Comprehensive educational assessment for the states: The duplex design. Evaluation Comment (November 1987, pp. 1-16). Los Angeles, CA: Center for Research on Evaluation, Standards and Student Testing, UCLA.

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

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995). Generalizability

analysis for educational assessments. Evaluation Comment (Summer 1995, whole issue). Los Angeles, CA: Center for Research on Evaluation, Standards and Student Testing, UCLA.

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 (pp. 59-78; Report No. NCES 96-802). 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.

Shoemaker, D. M. (1971). Principles and Procedures of Multiple Matrix Sampling (Technical Report 34). Inglewood, CA: Southwest Regional Laboratory for Educational Research and Development.

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: Costs of Matrix Sampling of Test Items. ERIC Digest.
Note: For a related discussion of matrix sampling, see ERIC Digest EDO-TM-03-08 TM 035 219).
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: Costs, Matrices, Reliability, Sampling, Scoring, Test Construction, Test Items, Validity
Identifiers: ERIC Digests