Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 8, Number 16, July, 2003 ISSN=1531-7714

Matrix Sampling of Items in Large-Scale Assessments

Ruth A. Childs
Ontario Institute for Studies in Education of the University of Toronto

Andrew P. Jaciw
Stanford University

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 both 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 paper, matrixed designs are compared with more traditional designs in nine categories of costs: development costs, materials costs, administration costs, educational costs, scoring costs, reliability costs, comparability costs, validity costs, and reporting costs. In choosing among test designs, a testing program should examine the costs in light of its mandate(s), the content of the tests, and the financial resources available, among other considerations.

Imagine that you must create a test of science knowledge and skills to be administered to all Grade 5 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 Grade 5 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 Grade 5 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?

Matrix Sampling of Items

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. 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. Consequently, this paper will focus only on sampling of items.

For example, 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. Figures 1 and 2 illustrate designs in which all students take the same number of items. Figure 3 shows a matrixed design.

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 be "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. Figure 4 illustrates this design.

Figure 1: Option 1- All Items Are Administered to All Students

Student	Constructed-Response Items (N Items = 32)	Multiple-Choice Items (N Items = 16)
1	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	xxxxxxxxxxxxxxxx
2	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	xxxxxxxxxxxxxxxx

Figure 2: Option 2 - The Same Selected Items Are Administered to All Students

Student	Constructed-Response Items (N Items = 8)	Multiple-Choice Items (N Items = 4)
1	xxxxxxxx	xxxx
2	xxxxxxxx	xxxx

Figure 3: Option 3 - A Matrix Sampling Design

	Constructed-Response Items (N Items = 32)				Multiple-Choice Items (N Items = 16)			
Student	Set 1	Set 2	Set 3	Set 4	Set 1	Set 2	Set 3	Set 4
1	xxxxxxxx				xxxx			
2		xxxxxxxx				xxxx		

3	xxxxxxxx	xxxx
4	xxxxxxxx	xxxx

Figure 4: Option 4 - A Partial Matrix Sampling Design

	Constructed-Response Items (N Items = 32)								Multiple-Choice Items (N Items = 16)							
		Matrix Set								Matrix Set						
Student	Common	1	2	3	4	5	6	7	Common	1	2	3	4	5	6	7
1	xxxx	xxxx							xx	xx						
2	xxxx		xxxx						xx		xx					
3	xxxx			xxxx					xx			xx				
4	xxxx				xxxx				xx				xx			
5	xxxx					xxxx			xx					xx		
6	xxxx						xxxx		xx						xx	
7	xxxx							xxxx	xx							xx

Jurisdictions Using Matrix Sampling

Several states, including Kentucky, Maine, Maryland, Massachusetts, Oregon, Pennsylvania, and Wyoming, have used partial matrix sampling in their testing programs. For the 1994/1995 school year, for example, the Kentucky Instructional Results Information System (KIRIS) assessed students in grades 4, 8, and 11 in six subject areas: Reading, Math, Science, Social Studies, Arts & Humanities, and Practical Living/Vocational Studies (Kentucky Department of Education, 1997). In order to adequately sample the content in each subject area, 29 items (grade 4) or 30 items (grade 8) were included in the assessment for Reading, Mathematics, Science, and Social Science and 24 items each in Arts & Humanities and Practical Living/Vocational Studies. The open-response items were intended to require 10 minutes each to complete, so that administering all the items to each student would require four to five hours of testing time for each subject area. Across the six subject areas, each student would be required to spend about 28 hours in testing, if required to take all the items.

To limit the testing time for each student to one 90-minute testing period for each subject area (additional time was provided to students who needed more than 90 minutes to complete the items), partial matrix sampling was employed. In Reading, Mathematics, Science, and Social Studies, five or six items in each subject area were selected to be common items, while the remaining items are divided across the various forms. (In contrast, in the Arts & Humanities and Practical Living/Vocational Studies areas, all items are matrix-sampled.) On the grade 8 mathematics test, for example, six items were selected as common items, to be administered to every student. The remaining 24 items (out of the total of 30 items that were selected to provide coverage of the subject area) were divided into groups of two. Each test form contained the six common items, plus one group of two additional, matrix-sampled items.

The 1999 *Massachusetts Comprehensive Assessment System* (MCAS) used a similar design. Approximately eighty percent of the items on the MCAS were common items, with the remaining twenty percent matrix-sampled (Massachusetts Department of Education, 2000). Both the KIRIS and the MCAS have relied on the common items as the basis for individual student scores. The Wyoming Comprehensive Assessment System, although using a similar test design, bases student scores on both common and matrixed items.

The National Assessment of Educational Progress (NAEP) uses a matrixed design in which all the items are divided into blocks, which are then combined into booklets in such a way that every block is paired once with every other block in one booklet. This design minimizes the testing burden for individual students, but still permits the estimation of proficiency distributions for large groups of students – especially at the national or state levels. NAEP does not yield results for individual students, however.

Similarly, the Dutch National Assessment Program (DNAP) has used a matrixed design in which a large number of items are developed, then divided into "instruments." A sample of elementary school students in the Netherlands is selected and each student receives only a sample of the instruments. Like the NAEP, the DNAP yields only aggregate results.

Costs of Matrix Sampling

In the scenario with which we began, we identified two of the issues that must be considered when deciding what design to use in a testing program: content coverage and testing time. Additional considerations include printing and scoring costs and the precision of student- and group-level scores. These considerations can be thought of as different types of costs:

- Development costs: Time and money expended in developing (writing, editing, reviewing, and pilot and field testing) new items
- Materials costs: Time and money expended in printing and shipping test booklets and other materials
- Administration costs: Time expended by teachers and other school personnel related to administering the test
- Educational costs: Time expended by students in preparing for and taking the test; changes in teachers' classroom practices and coverage of the curriculum because of the test; changes in schools' allocation of resources because of the test
- Scoring costs: Time and money expended in scoring the tests, either electronically or with trained judges
- Reliability costs: Changes in accuracy and consistency of test scores
- Comparability costs: Changes in the extent to which different students' test scores can be compared
- Validity costs: Changes in how well the test scores reflect the construct the test is intended to measure
- Reporting costs: Changes in how easily test scores can be explained to teachers, parents, and the general public

These costs are ones that, in our experience, testing programs commonly consider (although not every program considers every cost). It is possible, however, that some programs may have additional costs not included in this list.

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. To illustrate how these costs affect testing programs, in the following pages, each cost is described in general and then with respect to the four hypothetical design options for the science test. (Refer to Figures 1-4 for illustrations of the four design options.)

Development 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, more participation by schools, and so forth. In small jurisdictions, developing large numbers of items may be particularly burdensome for two reasons: (1) the cost of developing additional items raises the per student cost of testing more quickly when there are fewer students who will be taking the test; and (2) the numbers of schools and students available to pilot and field test new items are limited.

In general, developing more items requires more resources and is therefore more costly. Of the four hypothetical design options, Option 2 has an advantage over the other options in terms of development costs in that only a few items need be developed, with all of the students responding to this limited number of items. Each of the other options requires that a larger pool of items be developed, although each utilizes the larger number of items in a different way.

Materials Costs

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, cost more to ship. 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 multiple versions are more expensive to produce than a single version of a test.

If a test is to be administered by computer instead of using paper and pencil, 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.

In addition to test booklets, other materials must be printed and mailed to the schools and to parents. For schools, these include 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, because some parents do not have Internet access, printed materials are still required.

Options 1 and 2 have the advantage with respect to materials costs because they require the production of only a single form.

Administration Costs

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, paper and pencil, and only a single version is to be administered, the time to prepare for the administration may be minimal. However, if the test is longer, students must perform tasks in addition to writing responses, and/or 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.

Options 1 and 2 also have the advantage with respect to administration costs because they involve the use of only a single form, and are therefore simple to administer. The use of multiple forms, as is the case for Options 3 and 4, can sometimes require that the forms be distributed among students in a prearranged way. This increases the administration burden and introduces an additional source of potential error.

Educational Costs

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.

The options vary in educational costs. Option 1, which requires all students to take a long test, will require more testing time; however, by covering a broader range of the curriculum, it will minimize the possibility that teachers narrow their instruction to what they perceive to be the material that will be covered on the test. Option 2 requires all students to respond to the same small set of items and so is not costly in terms of administration time. However, by covering only a small sample of the curriculum, it may encourage teachers to focus disproportionately on those parts of the curriculum. This could happen, for example, if teachers assumed that the content of the items used in the previous year's test was particularly important and likely to be addressed again in future assessments. Both Options 3 and 4, by being matrixed in whole or in part, keep testing time short, and suggest that more of the curriculum is important.

Scoring Costs

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 both time-consuming and expensive. As the amount of student work to be marked increases, the amount of scoring, of course, increases. It is not so obvious, though, that 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.

Option 1 will be more costly than the other designs in terms of the total number of responses that need to be scored because every student answers every item in the pool. Options 1, 3 and 4 require training judges to score a wider range of items than are scored in Option 2. This additional training translates into higher scoring costs.

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 it placing students in one of several performance levels or yielding 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 no necessarily improve score reliability for students, schools, and districts. To improve score reliability in an IRT analysis 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.

Option 1, in which all students respond to all items in a large pool of items, ensures reliability of scores at the student level by including many items. This also helps to ensure that school-level scores are reliable. The use of a smaller number of items in Option 2 decreases the reliability of both student- and school-level scores. Options 3 and 4, by including matrixed items, greatly expand the number of items administered at the school level and therefore result in more reliable school-level scores. A drawback to using these designs is that individual students respond to smaller sets of items and so student scores are less reliable.

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 (IRT), 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 effect comparability. IRT models require at least several hundred responses per item. Some testing programs, such as the one in Alberta, have teachers grade their students' work and record the results for inclusion in class marks before sending the tests off to the state or provincial testing agency. IRT analyses are not possible in such situations, so that if students in the same class respond to different sets of items it will be difficult for teachers to estimate comparable scores.

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, Linn, Brennan, & Haertel, 1995; Fitzpatrick, Lee, & Gao, 2001; Haertel & Linn, 1996; Mislevy, Beaton, Kaplan, & Sheehan, 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, Linn, Brennan, and Haertel (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.

For ease of comparability, Options 1 and 2 have the advantage because all students respond to the same items. In designs of Options 3 and 4, it may be more difficult to compare student scores.

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 a student is able to demonstrate his or her knowledge on the test. For example, test instructions that are very confusing 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 that are important but do not affect the four design options that are used as examples in this paper, include a test's reading level or the inclusion of extraneous concepts that may be less familiar to some students than to others.

Options 1, 3 and 4 cover more of the curriculum than does Option 2 and therefore yield more valid results at the school level, if the purpose of the test is to measure mastery of the curriculum. Options 2, 3 and 4 test students using fewer items than Option 1, and therefore have diminished validity at the student level. That is, with fewer items, less of the curriculum is addressed leaving open the possibility that for a given student only certain subdomains of the curriculum are being tested. Option 1 includes more items, which gives students a greater opportunity to demonstrate understanding of the construct being tested; however, increased test length may come at the price of student fatigue, which may undermine the validity of subsequent inferences.

Reporting Costs

A more complex test design may require more explanatory materials, 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.

Options 1 and 2 are the easiest to explain. The same items are administered to all students. Options 3 and 4 are more difficult to explain. Justifications may need to be provided for why different students receive different items.

Summary of Costs by Option

The main disadvantage of Option 1, in which all items are administered to all students, is the longer testing time: about six hours. Option 1 also requires the most items per student to be marked, increasing scoring costs. However, Option 1 has advantages too: (1) the larger number of items per student increases the reliability of student scores and increases curriculum coverage, and (2) the design is easy to explain.

Option 2, in which the same smaller set of items is administered to all students, has the smallest item development, materials, and scoring costs. In addition, the time required to administer the test is less than two hours, an advantage over Option 1. However, the reliability of the student scores suffers and the coverage of the curriculum is less. Like Option 1, Option 2 has the advantage of being easy to explain.

Option 3 is a fully matrixed test, in which each student receives one of four non-overlapping forms. This option requires the same item development costs as Option 1, but has larger materials and administration costs. This design combines the advantages of a shorter testing time with increased curriculum coverage. The reliability and validity of student scores are comparable to those for Option 2 and the reliability of school-level scores is better, but the comparability of student scores may suffer. In addition, Option 3 is more difficult to explain than Options 1 or 2.

Finally, Option 4 is a partially matrixed test, in which each student receives some items that are common across all students and some matrixed items. Option 4 requires the same item development as Option 1, but higher administration costs and the highest materials costs (because it requires the most booklet versions). However, Option 4 has shorter testing time than Option 1, while having the same curriculum coverage at the school level. The other

advantages and disadvantages are similar to those for Option 3, though student score comparability may be improved.

Is Matrix Sampling of Items a Viable Approach?

How should a state's or province's testing program that is considering using matrix sampling of items proceed? As outlined in the previous section, every test design, whether or not it involves a matrixed component, carries with it certain costs. The nine categories of 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.

If the mandate of a testing program involves reporting scores at both student and school levels, a design involving the partial matrix sampling of items may be a viable alternative. Referring back to the test designs that were considered, Option 1, though optimal in principle, may not be practical because of the time that would be required to administer the full set of the items. Option 2, by using a small number of items, leads to a serious compromise of validity at both student and school levels. Option 3, the fully matrixed option, is optimal for determining school-level but not student-level scores. The partially-matrixed design, Option 4, allows the reporting of both student and school-level scores. Obviously there are costs with this design, as well; however, on the whole it is both feasible and informative.

For any of the designs that administer a reduced number of items to each student (i.e., Options 2, 3, and 4), whether there is enough information at the student level to report subscores may be a concern. For example, research by Gao, Shavelson, and Baxter (1994) suggests that each student must answer at least nine or ten performance tasks to avoid very large effects of person-by-item interactions. To produce reliable subscores, even more items may have to be administered. Given that there are limits in test administration time, it may not be feasible to administer enough items to support student-level subscores. Instead, only overall scores might be reported at the student level, while both overall scores and subscores are reported at the school level. The partially matrixed design is especially well suited to this purpose. It improves on the fully matrixed design because it provides more comparable student-level scores, while still adequately sampling performance at the school level. It should be noted, however, that the items in the fully matrixed design will be administered more often in a given school than the matrixed items in the partially-matrixed design. Depending on the size of the school, this difference may affect the reliabilities of school-level subscores.

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.

Note

This article is based on a paper presented 31 January 2002 at the Symposium on Provincial Testing in Canadian Schools: Research, Policy & Practice, Victoria, BC, Canada.

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, University of California at Los Angeles.

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, University of California at Los Angeles.

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.

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.

Kentucky Department of Education. (1997). *KIRIS accountability cycle 2 technical manual.* Frankfurt, KY: Author.

Massachusetts Department of Education (2000). *The Massachusetts Comprehensive Assessment System: 1999 MCAS technical report*. Malden, MA: Author.

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.

Popham, W. J. (1993). Circumventing the high costs of authentic assessment. *Phi Delta Kappan, 7,* 470-473.

Shoemaker, D. M. (1971). *Principles and procedures of multiple matrix sampling* (Technical Rep. 34). Inglewood, CA: Southwest Regional Laboratory for Educational Research and Development.

Correspondence concerning this article should be addressed to Ruth A. Childs, OISE/UT, 252 Bloor Street West, 11th Floor, Toronto, Ontario M5S 1V6. E-mail: rchilds @oise.utoronto.ca

Descriptors: Sampling; Test Items; Item Sampling; Matrix Sampling; Research Design

Citation: Childs, Ruth A. & Andrew P. Jaciw (2003). Matrix sampling of items in large-scale assessments. *Practical Assessment, Research & Evaluation*, 8(16). Available online: http://PAREonline.net/getvn.asp?v=8&n=16.