Journal of Mathematics Education at Teachers College
Spring–Summer 2012, Volume 3



# A Population of Assessment Tasks

Phil Daro
University of California at Berkeley

Hugh Burkhardt
Shell Centre, University of Nottingham
University of California at Berkeley

We propose the development of a "population" of high-quality assessment tasks that cover the performance goals set out in the *Common Core State Standards for Mathematics*. The population will be published. Tests are drawn from this population as a structured random sample guided by a "balancing algorithm."

*Keywords*: assessment tasks, CCSSM, test item development.

Tests are not "just measurement." When tests are found to "steer the system," as they do when they dominate high-stakes consequences, their validity for steering must be evaluated. Indeed, the behavior of assessment clients counts as crucial validity evidence. We cannot blame those being judged for focusing their resources and effort on actions that they think will most efficiently increase scores, when those scores have so much influence on their and their students' lives. Therefore, the actions of educators in response to tests must be taken as evidence on the validity of using that kind of test for high-stakes accountability. We need more empirical studies of client interpretation and behavior in response to assessment. A test that might be just fine when used as an unobtrusive measure might equally be profoundly invalid when used as a high-stakes criterion.

Diane Ravitch, David Berliner, and others have warned us that the balance of learning activities in most classrooms is strongly influenced by the items in the state accountability tests that the students will face. The weeks of "test prep" in mathematics classrooms are evidence that teachers know this. There are various effects. There is reduced emphasis on subjects not tested, like science at most grades, art, social studies at most grades, and so on. There are effects on the length and types of problem assigned to students for class and homework, narrowing the range to those that appear on the state test. In short: *what you test is what you get*.

This effect means that high-stakes tests can drive the quality of teaching and learning up[1] or, as has been more usual with state accountability tests, down (see e.g. Wright, 2002; Fairtest, 2012). To design and deliver state tests that actively encourage improvement presents major challenges. Here we propose a new tool[2] that, if implemented, will help all involved achieve that important goal.

We propose the development of a "population" of high-quality assessment tasks that cover the performance goals set out in the *Common Core State Standards for Mathematics* (CCSSM).

Psychometric theory (see for example Mislevy et al., 2002) regards any well-constructed test as a sample of "items,"[3] selected randomly (usually stratified according to a "construct") from a "population" of items. This population of items represents the domain being measured. As with any sampling design, the inferences about the population depend on the relationship between the sample and the population. A population of random samples will have the same mean as the population of individuals. If the sample is not random, the inference is endangered in a number of well-known ways.

This population of items can be difficult to describe in relation to the knowledge and proficiencies that the test is attempting to measure. Often, an unwarranted epistemological assumption is made: that the knowledge and proficiencies can be segmented into "topics," the union of which is equivalent to the knowledge and proficiencies of interest. This assumption is wrong in the many cases where links across topics are important features of performance. In mathematics, for example, the ability to integrate mathematical concepts and skills with mathematical practices is at the heart of "doing mathematics." Similarly, in English Language Arts good

[1] For cases of positive impact, see, for example, Section 4 of Burkhardt (2009).

[2] R. D. Bock (1993, 1997) proposed something similar for NAEP. See also Bormuth (1970).

[3] While recognizing this general use of the term "item," we prefer the more descriptive "task" for any performance that students are asked to tackle in assessment, reserving "item" for any prompt that relates to a single score point (its original meaning, coming from statistics). For many short items, of course, the two are the same.

writing depends on choosing, and using effectively together, vocabulary, grammar, and syntax to communicate meaning.

A related fallacy arises when there is a lack of explicit criteria for deciding when a topic is adequately measured by a set of items classified as belonging to that topic. For example, in an art test the item "Is blue in the picture?" would be classified as belonging to the topic "What colors are in the picture?" But if the full set of items on this topic omitted any questions about colors other than primary colors, it would not be a measure of the respondent's knowledge of color. Criteria need to specify the balance, variety, and range of items that would satisfy the topic.

When the form of the test, and the types of performance to be tested, are well-established, well-understood, and unchanging, the test might be fair for high-stakes, even if what is measured has a peculiar relationship to what is valued. However, the US is now in a very different situation, with widespread acceptance that state tests must change substantially if they are going to assess the range of performances implied by the CCSSM. Assessment design has to accommodate more complex kinds of evidence (Mislevy & Haertel, 2007). The change is particularly stark in mathematics, where the mathematical practices involve substantial chains of reasoning that cannot be assessed through short items. This makes clear the inadequacy of most traditional assessments.

## The Proposal

Here we recommend that consideration be given to making clear in the most direct way possible the range, variety, and balance of the types of tasks that students may be tested on by:

- creating and making public the population of assessment tasks from which any test can be drawn as a sample; and
- describing the kind of procedure by which the samples for inclusion in a test will be constructed—here called a "balancing algorithm."

These two elements play different, complementary roles.

*The task population* is an enabling entity. It communicates in explicit form the range and variety of task types that span the performances that CCSSM identifies as important. Thus it offers a resource for test developers that can save them and their clients time and money. It will be particularly useful in areas of task design that are outside established expertise.[4]

*The balancing algorithm* plays a complementary role. There will be different views on the proportions of the various types of tasks that a balanced test should involve. These will be reflected in the balancing algorithm that is chosen by those responsible for the test—a normal part of test construction. The population simply offers to such bodies the full range of possibilities; their balancing algorithm will make explicit the value system of the community for which the test is to be designed—one that may be compared with the classification scheme included with the population.

The population together with the balancing algorithm used by the provider of their students' tests offers a resource to teachers, exemplifying the range of performances in which their students should gain experience.

## On What Principles Will a Population of Tasks Be Based?

To meet these goals, the population needs to be:

- *comprehensive*, covering all the main types of tasks within the learning and performance goals of the curriculum—in this case, those set out in the *Common Core State Standards* and the consortium documents based on them;
- *large enough*, for each type of task, so that reliably *learning the solutions* to all the tasks of that type will be so burdensome that attempting it will clearly be a poor teaching strategy, compared to *learning how to solve* such problems;[5] and
- *structured* in ways that support the design of the balancing algorithm(s) that will be used for constructing tests.

There is by now a substantial "literature" *of* assessment tasks of many different types; the population will be based on, and initially drawn from, this literature. Equally, there is a parallel literature *about* such tasks and their classification, which will guide the structuring.

## How Can Such a Population Be Constructed?

First, the big picture. Finding, sifting and sorting, and assembling the population from the task literature is a substantial challenge—but a challenge of the kind and scale of a typical research project. A pilot version could be delivered within a year, with another year or two for refinement. A mechanism for maintaining and refreshing the population can then be developed, as indeed it must be. This must be a "learning field," with contributions from both the mathematics education and psychometric

[4] It has not escaped our attention that this tool will have other potential uses—for example, in research including the evaluation of curricula or other improvement programs, where the instruments used (often without question) have usually been too narrow to address the principal learning goals of ambitious programs like the NSF mathematics curricula—or CCSSM.

[5] But if a student learned them all, we would surely be happy!

communities creating better examples and new types of tasks to address unmet learning needs.

*Finding Candidate Tasks*

The principle here is to cast the net as widely as possible. Designing rich assessment tasks that allow all students *to show what they know, understand, and can do*[6] across the range of practices and content set out in CCSSM is among the most challenging areas of educational design. The objective is to find for the population the products of outstanding task designers from around the world. Informal approaches suggest that, with appropriate recognition, they will be delighted to contribute. Equally, reactions from assessment providers suggest that most will be happy to offer tasks to a well-run scheme that will provide some financial return. Contributors will, like everyone else, have access to the population for their own test construction.

We want the population to include tasks that offer students and their teachers something of the pleasure that doing mathematics well can provide. For this we need to liberate the imaginations of task designers—and those who will be selecting tasks.

*Establishing Acceptance Criteria*

As always, the challenge is to balance rigor with inclusiveness. On the one hand, if *any* task can be included, the user is faced with all the costs of development that inclusion in a serious test will require; on the other, if a full analysis of each task's psychometric properties based on trials with a representative student population is demanded, far too few rich tasks are likely to qualify, at least initially.

A sensible balance might be based on two principles: *inclusiveness* and *transparency*—

- tasks with some evidence that they work well with students should be accepted, provided they come with
- information on the type and extent of trialing already undertaken and its outcomes.

For example, one might require at least:

- the task prompt,
- one or more scoring rubrics, and
- 10 samples of student work on the task from trialing, unscored and scored, illustrating its accessibility to students of all kinds
- *along with* a summary of the trialing process involved (source and number of students, their ability range, circumstances of performance,…).

Users can then judge what further development they feel is necessary, and estimate its cost.

*Establishing Licensing Procedures*

Many of the tasks that one will wish to include in the population will have copyright or other entailments. Licensing arrangements will be developed and offered to those controlling the intellectual property rights. These will involve:

- free non-commercial use, e.g. by teachers in their own classrooms or professional development leaders in activities provided free by the school system; and
- payments to the IP owners for commercial use—for example, by test developers for the tasks used (with a fee to the population managers).

While providing income for the rights holders, the cost to a user will be substantially less than the design and development cost of a comparable task.

*Exclusivity:* Exclusivity is an issue that needs to be addressed. Test providers are accustomed to exclusive use of the items in their tests (even when there is nothing unusual about them). This partly reflects the investment in establishing their psychometric properties. It is also part of the long-running "secure test" phenomenon. However, the essence of the population is that it is open. Further, rich tasks are memorable so that no test involving them, once used, is really secure.[7] A reasonable solution might be to ensure that selected tasks were not also licensed to another test provider[8] until the test has been used.

*Security:* The essential element in test security is simply that *no-one preparing for the test knows which tasks will be in it*—a responsibility that test providers are used to, that remains straightforward to achieve when using the population.

*Classifying and Sorting*

This is an area that is sure to produce alternative views, since individuals and groups have been generating their own schemes for task classification for a very long time. We will discuss this below; for the moment we make just two points:

- Many classification schemes can be accommodated.
- It is usually possible to relate two schemes with sufficient precision for the purpose of test

[6] Our goal in the words of the influential UK Cockcroft Report (1982).

[7] It is widely recognized that the concept of a "secure test" is a social construct rather than a realization of what the words imply. Examination providers outside the US rarely rely on a test remaining secure after the date of the examination. "Past papers" are widely used in instruction—excellent, provided the tasks are good.

[8] Such tasks could also be withdrawn from the population but, in the age of efficient searching, this might make them easier to identify.

balancing—never a precise matter, given the many dimensions of performance in mathematics.

So if a user wants to use their own scheme, rather than one of those provided by the population database, the user is free to do so.

*Governance*

While the creation of the population is a job for a project team, its longer term value requires a governance structure that should be contemplated from the start. This might involve authoritative governance or more open source structures. In view of the central importance of test tasks in the implemented curriculum in most classrooms, we suggest that the governance should be in the hands of distinguished mathematics educators, national and international, along with some mathematicians, psychometricians, and test providers. This group will have oversight over both the task collection and the classification schemes.

A plan of work for realizing these elements will be the subject for a more specific proposal, when reactions to this outline have been absorbed.

## Task Selection and Test Assembly

This is the key process in determining the quality of a test. While this is a responsibility of the test provider, the main goal of the "population" is to help those who produce tests achieve high quality. Here we shall not attempt to review what this means (see, for example, ISDDE, 2012[9]) but confine ourselves to making some key points in achieving high quality in a test, viewed as an assessment of mathematics as described, for example, in the *Common Core State Standards*:

*Validity is the priority.* We need tests of *high validity*, *adequate reliability,* and *reasonable efficiency* in the use of assessment time. In the past these priorities have often been reversed, so we have had very efficient, reliable tests—but only of fragments of mathematics at the "novice" level[10] of expertise that short items represent.

*Separate the stages of test development.* The ISDDE paper usefully identifies three aspects of providing a test: assembling an adequate pool of tasks; selecting the tasks to be included on a test from this pool; and handling the delivery of the test through to production of reports. These three functions have traditionally all been handled by the company providing the test; however, they involve very different skills, resources, and responsibilities. We agree with the ISDDE paper's suggestion that they be considered separately. This paper describes how the first function may be handled through the development and maintenance of a population, largely based on the existing literature of tasks. We believe that the second should be carried through by the body responsible for the test (a consortium, an individual state, or other body) rather than delegated to the vendor that will handle the third function of test delivery. Why do we suggest this?

*Task selection for the test is the key.* The variety and balance of task types in the test will largely determine the range and balance of what teachers in most classrooms will teach and their students will learn. (See Figure 1.) Getting this balance right is thus a major responsibility—in short, to produce "tests worth teaching to." Such responsibilities should not be delegated. The considerable challenge of ensuring that the test will be a balanced reflection of the learning and performance goals (now those set out in CCSSM) with adequate psychometric properties requires a wide range of expertise, particularly in the subject discipline. This requires a "Mathematics Board"—the kind of expert group we have described above for the related function of guiding the work on the population. Responsible bodies are well-versed in assembling such expert committees to guide their decisions.

## Task Classification and Population Structure

There is a multitude of ways to classify mathematical tasks. The good ones have much in common. Choosing one is ultimately a matter for the Population Expert Group and, if they are not happy with that choice, for the Mathematics Boards of test providers. Here we shall point out important factors in choosing or designing such a scheme, illustrating them with the model shown in Figure 1.

Again this is an area where CCSSM, reflecting international standards in mathematics education, requires much more than is offered by current practice. Traditionally, attention has focused exclusively on the separate concepts and skills that are tested by short items—indeed, that is all that has been assessed. More recently, some other factors have been addressed in simple ways—for example, "depth of thinking" on a 3-point scale. CCSSM points out that doing mathematics involves

[9] The report of the 2010 Working Group of ISDDE, the International Society for Design and Development in Education, on the design of examinations in support of policy.

[10] It is the essence of psychometrics that its methods are independent of the content of what is assessed so it is not surprising that priorities are different from those of subject specialists. Psychometric ambition has often gone well beyond the needs of users. For accountability purposes, one reliable mathematics score per student is the most that is required. This does not require, for example, that all items and students are put on the same IRT scale. Consistency between scores from complete "parallel tests" is all that is needed.

## Dimensions of Balance

*Mathematical Content Dimension*

Mathematical **content** in each task will include some of:

- **Number and Operations** including: number concepts, representations relationships and number systems; operations; computation and estimation.
- **Algebra** including: patterns and generalization, relations and functions; functional relationships (including ratio and proportion); verbal, graphical tabular representation; symbolic representation; modeling and change.
- **Measurement** including: measurable attributes and units; techniques tolls and formulas.
- **Data Analysis and Probability** including: formulating questions, collecting, organizing, representing and displaying relevant data; statistical methods; inference and prediction; probability concepts and models.
- **Geometry** including: shape, properties of shapes, relationships; spatial representation, location, and movement; transformation and symmetry; visualization, spatial reasoning, and modeling to solve problems.

*Mathematical Process Dimension*

**Phases** of problem solving include some or all of:

- Modeling and Formulating
- Transforming and Manipulating
- Inferring and Drawing Conclusions
- Checking and Evaluating
- Reporting

**Processes** of problem solving, reasoning and proof, representation, connections, and communication, together with the above phases, will all be sampled.

*Task Type Dimensions*

- **Task Type** will be one of: design; plan; evaluation and recommendation; review and critique; non-routine problem; open investigation; re-presentation of information; practical estimation; definition of concept; technical exercise.
- **Non-routineness** in: context; mathematical aspects or results; mathematical connections.
- **Openness** – tasks may be: closed; open middle; open end with open questions.
- **Type of Goal** is one of: pure mathematics; illustrative application of the mathematics; applied power over a practical situation.
- **Reasoning Length** is the expected time for the longest section of the task.

*Circumstances of Performance Dimensions*

- **Task Length:** In these tests most tasks are in the range 5 to 15 minutes, supplemented with some short routine exercise items.
- **Modes of Presentation, Working and Response:** These tests will be written.

*Construct Irrelevant Difficulties*

- **Inconsiderate text:** language of text is inconsiderate to the reader, presenting awkward syntax, ambiguous references, ego-centric assumptions, distracting terminology or other difficulties that will interfere with assessment of construct. Such inconsiderate text may suppress performance of students with language processing issues.
- **Item presentation conventions:** conventions in diagrams, labeling, and language that belong to test formats or print formats generally rather than to mathematics itself.

From *Balanced Assessment for the Mathematics Curriculum*, an NSF-funded project (see Balanced Assessment, 1997-99)

Figure 1.

mathematical practices that represent deeper connected thinking of different kinds, all requiring longer chains of reasoning.[11]

In any design, there is a trade-off between competing "goods." Here we list some factors that are "goods" in any classification scheme:

*Low-inference interpretation*. It is obviously desirable that the classification scheme is clear and unambiguous, in the objective sense that different people with some expertise in mathematics education will classify a given task in much the same way. This is easier if the factors used are close to directly observable rather than, for example, depending on inference from a deep theoretical model.

*Multi-dimensionality*. One of the big changes needed is in recognizing the many dimensions of performance that doing mathematics involves. Any model should consider the various dimensions shown in the table, excluding any of them (or their near-equivalents, e.g. "practices" for "processes") only for good reasons. (In contrast, typical current short item tests sample across only one of these dimensions—that of "content." In terms of the dimensions of balance in the table, they are all *closed routine technical exercises* in *pure math*, involving only *transforming/manipulating* with *reasoning lengths* of less than 2 minutes—thus they are inevitably at what the ISDDE paper calls "novice level," assessing knowledge of the "tools of the trade" but not the ability to use them purposefully.)

*Holistic as well as analytic dimensions*. If one focuses only on the separate elements of mathematics that a task involves, it is easy to miss the point of the task. The holistic "task type" dimension has proven invaluable in classifying tasks. The list of types in the table can surely be improved but the principle is sound—and important.

*Simplicity and ease of use*. Simplicity and ease of use is a good feature of any design—provided it fulfils its function. Experience suggests that, as it is developed, the population of tasks will develop more than one classification scheme. A simple one might just include:

- Grade range of suitability;
- Task type – a fuller list than in the table, with clear links to (non-)routineness;
- Content areas at two levels;
- Task length and Reasoning length; and
- Level of expertise required ("novice," "apprentice," "expert").

A more sophisticated scheme will, at least, give more detail on the *mathematical practices* and *problem solving phases* involved.

[11] For example, three of the four "claims" in the framework adopted by the SMARTER Balance Assessment Consortium address: solving well-posed problems; constructing and critiquing reasoning; modeling (see SBAC, 2011).

## Conclusions and Next Steps

This paper sets out a new approach to the implementation of assessment, including high-stakes tests, that can serve instruction and learning, and put accountability on a valid basis. It is written from our perspective as mathematics educators rather than psychometricians. While the tools of psychometrics focus on the statistical properties of assessment, standard scientific methods teach us that systematic errors must equally be taken into account. Minimising these in assessment means collecting the right balance of evidence. Assessing the wrong things, however accurately, is misleading—and can be damaging.

## Acknowledgements

We have learnt a great deal from conversations with many people including test designers and providers in the US, the UK, and Australia. We are particularly grateful to members of the Mathematics Assessment Resource Service (MARS) and to Bob Mislevy for comments on an earlier draft.

## Notes

We invite comments and suggestions on this paper and, from holders of excellent mathematical tasks, in-principle offers to contribute to the population.

## References

Balanced Assessment. (1997–99). Balanced Assessment for the mathematics curriculum, eight volumes of classroom assessment. Parsippany, NJ: Dale Seymour Publications, Pearson Learning.

Bock, R.D. (1993). Domain referenced reporting in large scale educational assessments. Paper commissioned by the National Academy of Education for the Capstone Report of the NAE Technical Review Panel on State/NAEP Assessment.

Bock, R.D., Thissen, D., & Zimowski, M.E. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, *34*, 197–211.

Bormuth, J.R. (1970). On the theory of achievement test items. Chicago: University of Chicago Press.

Burkhardt, H. (2009). On strategic design. *Educational Designer*, *1*(3). Retrieved from: http://www.educationaldesigner.org/ed/volume1/issue3/article9/

Cockcroft Report. (1982). *Mathematics counts*. Report to the UK Government of the committee of enquiry. London, UK: HMSO.

Fairtest. (2012). The case against high-stakes testing. Retrieved from http://fairtest.org/arn/caseagainst.html

ISDDE. (2012). *High-stakes examinations to support policy: Design, development and implementation*. Paul Black, Hugh Burkhardt, Ian Jones, Glenda Lappan, Daniel Pead, and Max Stephens, for the ISDDE 2010 Working Group on Examinations and Policy, to be published in *Educational Designer*; retrieved from http://www.publications.parliament.uk/pa/cm201012/cmselect/cmeduc/writev/1671/exb52.htm

Mislevy, R.J., Wilson, M.R., Ercikan, K., & Chudowsky, N. (2001). Psychometric principles in student assessment. In D. Stufflebeam & T. Kellaghan (Eds.), *International handbook of educational evaluation*. Dordrecht, the Netherlands: Kluwer Academic Press.

Mislevy, R.J. and Haertel, G.D. (2007). Implications of evidence-centered design for educational testing. *Educational measurement: Issues & practice*. Retrieved from http://www.em2007.mpg.de/files/MislevyHaertel.pdf

SBAC. (2011). Draft content specifications for the summative assessment of the *Common Core State Standards for Mathematics*. SMARTER Balanced Assessment Consortium. Retrieved from http://www.k12.wa.us/SMARTER/ContentSpecs/MathContentSpecifications.pdf

Wright, W.E. (2002). The effects of high-stakes testing in an inner-city elementary school: The curriculum, the teachers, and the English language learners. *Current Issues in Education* [On-line], *5*(5). Retrieved from http://cie.ed.asu.edu/volume5/number5/