Balanced Assessment in Mathematics: the tests

MARS – Mathematics Assessment Resource Service
CTB McGraw Hill and MARS

Summary

These tests are designed for the many school systems that are seeking to improve the teaching and learning of mathematics in their schools by introducing goals and curricula based on national and international standards. The collection include 40-minute tests at each grade from 3 to 10, built from 5 to 10 minute tasks, which sample the broad domain of mathematical performance that national and many state standards specify. NCLB requires state tests to align with standards. The tasks demand substantial chains of reasoning and non-routine problem solving, covering the content and process areas specified in *Principles and Standards for School Mathematics*. Materials are provided to support reliable local scoring, which is also a contribution to teacher professional development. A scoring and reporting service is available, as are practice tests.

Tool Description

There are tests for each grade, currently Grades 3 through 10. Various alternative forms of the tests are now available from CTB and/or MARS.

- **Secure Tests** are published by CTB. There are a parallel forms for each grade.
- **Annual Tests** A new test form for each grade is developed by MARS each year, particularly to meet the needs of those systems that wish to use assessment as an active element in their professional development and school improvement program in mathematics. (*Assessment-led improvement: the MAC model* describes an outstanding example) Since every year's tests are different, teachers can review and discuss both tasks and responses with their students after they have taken the tests – a valuable learning activity that will directly help students to improve their performance. MARS supplies these tests to school systems under contract.
- **Practice Tests**, obtainable from CTB, are primarily designed to help teachers prepare their students for the tests. The two 40-minute forms for each grade, with scoring guidance, are also valuable in helping new users to a fuller understanding of this standards-based assessment – first by inspection, then in the classroom.
- **Balanced Assessment tasks** may also be included in CTB contracts for state or district tests, so as to improve the depth of knowledge, range and balance of the assessment.
- **Scoring Guides** are developed for each test.

Some task examples from across the grade range are shown below, along with an example of the scoring.

Background

Since the late 1980s there has been a coherent program of systematic development of standards, curriculum materials and assessment in mathematics education based on international standards. *Principles and Standards for School Mathematics* (PSSM), developed by the National Council of Teachers of Mathematics in consultation with the major societies of US mathematicians, is the first component. The National Science Foundation then funded

the development of a number of complete standards-based curricula for each grade, and assessment, each aligned with these standards. Balanced Assessment in Mathematics is a product of this process.

Design principles

These 40-minute tests are built from 5 to 10 minute tasks, which sample the broad domain of mathematical performance that the standards specify. The emphasis is on assessing students' performance on worthwhile tasks set in practical contexts. They demand substantial chains of reasoning and non-routine problem solving, addressing all the content and process areas specified in *PSSM.* The tests are based on the style of assessment used in most other countries and, now, in many school districts across the United States.

This assessment is designed by MARS, the Mathematics Assessment Resource Service, under the direction of a Mathematics Board, which includes teachers and recognized US and international experts in mathematics education and assessment. MARS is a US-based international team, which designs assessment for client states and school districts across the country. These tests build on this experience.

Scoring

- **Scoring.** While a full scoring service for the Secure Tests is available from CTB, all these tests are designed for local scoring. Materials and training to support the scoring are developed by MARS. The materials for scoring are presented in the Scoring Guide for each set of tests. They include point-scoring rubrics for every task, supported by unscored and scored examples of student work across the performance range.

  Live scoring training, for system leadership and/or each system's scorer teams, is designed to facilitate consistent and accurate scoring. It also represents invaluable and motivating professional development for the teachers involved.

  Linkage to national standards is provided through the cut-scores that come from the annual standards-setting procedure.

- **Standard-setting.** For the annual tests, cut scores for the four-level reporting scheme are set each year by the Mathematics Board. Their holistic professional judgments on samples of student work around each borderline between levels is informed by analysis of the tasks in relation to the standards, and by both qualitative and statistical information on student performance.

  For the secure tests, a sophisticated statistical 'anchoring' system that provides year-to-year and grade-to-grade scaling is available for some grades.

- **Reporting.** Reporting is designed to fit the needs of each system for various kinds of information. Accountability usually needs simple information – the overall level in mathematics attained by each student, based on his or her total score. Profile reporting gives summary information on different aspects of performance, in this case 5 content and 5 process headings – this offers rough guidance on areas of weakness that will merit more attention. Returning the scored papers to the classroom provides the fullest information for teachers and students.

  MARS offers to school systems a more detailed report on students' performance, linked to the Standards. These reports go beyond the statistics to give an educational analysis of what the assessment revealed about the students.

Professional Development

MARS also provides support for associated professional development, for and through balanced assessment of various kinds. Discussing student work on substantial mathematical problems brings out the major issues (of what is mathematics, of curriculum, and of instruction) in a down-to-earth practical way that teachers (and, through them, their students) readily understand.

Scoring training activities, as well as enabling systems to achieve reliable scoring, provides valuable and enjoyable professional development for teachers and others involved.

This contributes broadly and fundamentally to system improvement, as part of a broader range of professional development support for *professional development for and through performance assessment*, available from MARS.

Evaluative evidence

The tests have been used over many years by US school systems across the country. Feedback from users on use in assessment, in classrooms, and in professional development has been enthusiastic. There is further evidence that Balanced Assessment assesses a broader range of performance than state tests, and is comparably challenging overall. Ridgway, Crust, Burkhardt, Wilcox, Fisher, and Foster (2000) compared students' performance at grades 3, 5, and 7 on a standardized high-stakes, skills-oriented test (the California State STAR test) with their performance on a much broader standards-based test (a Balanced Assessment test)[1].

Scores on each test were divided into two simple categories: "proficient" or "not proficient." From 70-75% of the students at each grade level scored equivalently (either proficient or not proficient) on both tests. However, fewer than 5% of the students scored proficient on standards-based test and not proficient on the skills-oriented test, while more than 20% of the students were deemed proficient on the skills-oriented test but not proficient of the standards-based test.

That latter group of students, nearly 1/4 of the student population, was deemed "proficient" by the State on the basis of the STAR test, but only because of the narrowness of the test. Those students' low scores on the Balanced Assessment tests suggest that the "proficient" ratings on the STAR tests may be "false positives." That is, the students' proficiency is called into question when deeper and broader measures reflecting contemporary research are employed.

Strengths of this tool

The tests

- exemplify performance in mathematics
- recognize the broad range of aspects of performance that the standards demand
- enable states with high-quality standards to meet NCLB alignment requirements
- improve the influence of high-stakes assessment on the implemented curriculum

Teaching for the tests and scoring training both advance the professional development of teachers, leading to better understanding of the meaning of the standards. They provide essential support to other aspects of any *standards-based improvement* program, and help to increase student motivation.

Likely challenges

[1] Ridgway, J., Crust, R., Burkhardt, H., Wilcox, S., Fisher, L., and Foster, D. (2000). *MARS Report on the 2000 Tests.* Robert E. Noyce Foundation, San Jose, CA.

- Systems may be unwilling to introduce assessment that goes beyond state tests.
- There is a tradition, exacerbated by the under-funding of NCLB assessment requirements, of seeking inexpensive multiple-choice tests with little concern for what kind of mathematics they assess.

These pressures may be countered through the evidence that balanced assessment produces enhanced performance even on narrow state tests.

System administration may not provide teachers with enough professional development time to learn how to teach mathematics more effectively; however, professional development like this, directly linked to tests, is likely to be favored.

Availability

Further information on these tools can be found at and through:

www.ctb.com/bam/

www.balancedassessment.org

www.educ.msu.edu/MARS

Costs

About $2 per student test, plus local scoring costs including scoring training (1 or more sessions) and scoring time (10-20 student papers per scorer hour). For outside scoring and reporting service, about $10 per student test.

Design and development

Rita Crust, Hugh Burkhardt and others for *MARS*, partly funded by the National Science Foundation. Each test and its scoring rubrics are developed through two rounds of trials in invited US classrooms. Analysis of a standardizing sample of student responses to the live tests informs standard-setting. Scaling of some forms by CTB uses MARS 'Fat Anchor' procedure.

Exemplars

To illustrate the Balanced Assessment tests we provide on the following pages

- a few examples of tasks;
- an outline of the scoring approach.

Task exemplars

Three tasks are shown below. (Note: there is more space for working in the format of the actual tests.)

- **Languages** is a data analysis task involving specific representation – with a circle graph
- **Cube Shapes** is a problem involving generalization of a pattern
- **Gas** involves modeling a practical situation, particularly with the use of line graphs

Scoring

The task used to illustrate the scoring approach, **Magazine Cover**, involves geometrical shapes, co-ordinate systems, and communication.

All four tasks are non-routine in the sense that, apart from carrying through the procedures involved, deciding what to do is a significant part of the demand.

LANGUAGES

All 250 students at Holmes Junior High must take either French or Spanish for their foreign language requirement. Below are the number of boys and girls enrolled in French or Spanish.

	Number of Boys	**Number of Girls**
French	29	60
Spanish	121	40

1. Krishna was describing the Holmes Junior High students who take French. He said, "About one half of the students taking French are boys."
Alex said, "Wait a minute, it looks like about one third to me."
Who is correct? How do you know?

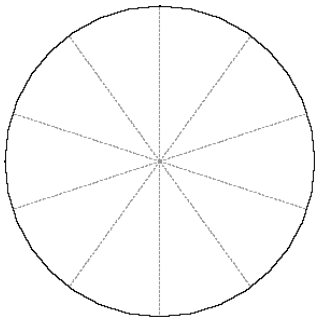
2. When a reporter for the school newspaper asked Ms. François to compare the popularity of French between boys and girls at Holmes Junior High, she said, "One fifth of the boys take French while three fifths of the girls take French."

Did Ms. François answer the question appropriately?
Why did she report her answer in fifths?

3. Ms. François was also asked to compare the popularity of French vs. Spanish across all of the students at Holmes Junior High.
Write a statement using percents that she could use to make this comparison.

4. The circle below is divided into 10 equal parts. Estimate, shade, and label the circle graph to represent the portion of all students at Holmes Junior High who are:

- boys taking Spanish,
- girls taking Spanish,
- boys taking French, and
- girls taking French.

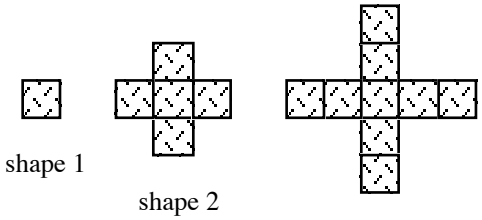


[10]

CUBE SHAPES

This problem gives you the chance to:
- find missing numbers in a pattern
- explain your thinking

Debbie makes shapes using wooden cubes.
The diagram below shows her first three shapes.
She uses one cube for shape 1 and five cubes for shape 2.



shape 1

shape 2

shape 3

shape 4

shape 5

(a) Draw diagrams showing shape 4 and shape 5 in the spaces above.

Debbie begins to make a table showing the number of cubes she needs to make each shape.

Shape number	1	2	3	4	5
Number of cubes	1	5	9		

(b) Find the number of cubes needed to make shape 4 and shape 5, and write these numbers in Debbie's table.

(c) How many cubes are needed to make shape 8? _____
Explain how you figured it out.

(d) Which shape number can Debbie make using 53 cubes?
Show how you figured it out.

(e) Find a rule or formula for figuring out the number of cubes needed to make shape ***n***.

[15]

GAS

This problem gives you the chance to:
- draw a graph to represent a real life situation
- show that you understand the concepts of rate and slope

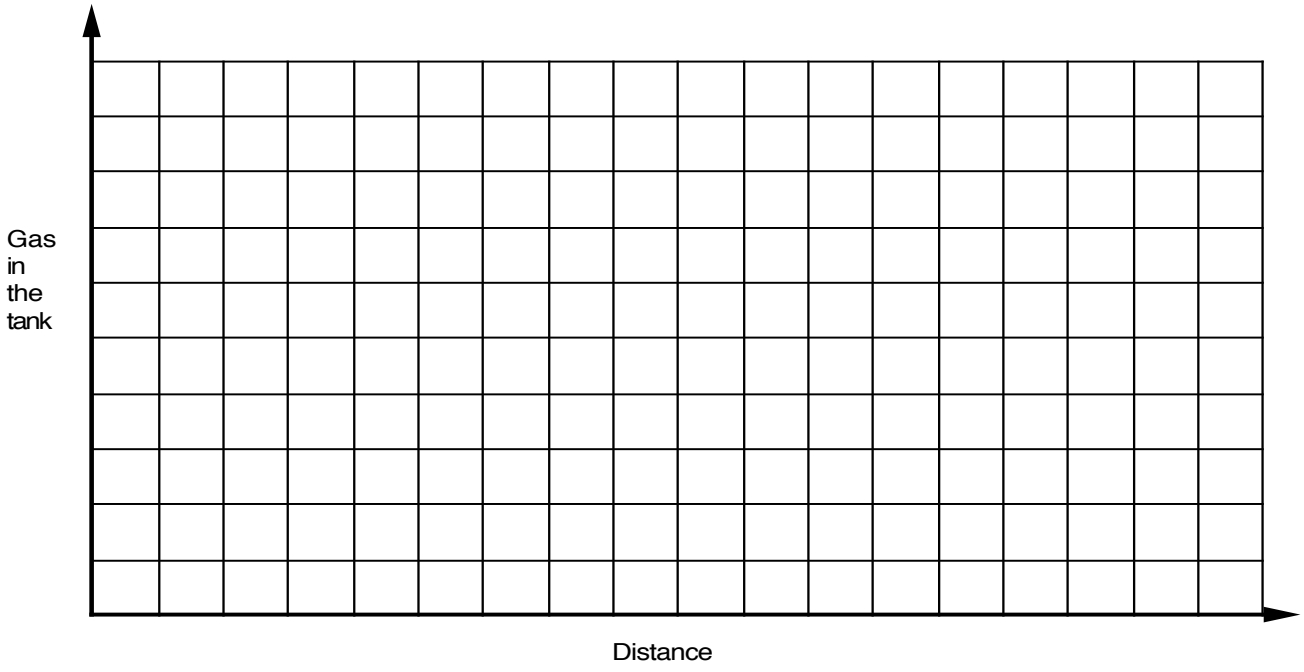
Read this story, then sketch a graph to show how the amount of gas in this car's tank might have varied during the trip.

Last weekend, I visited a friend in Los Angeles. Before I left home, I checked my gas. The tank was about one third full - so it must have contained about 3 gallons.

During my 200-mile drive down the freeway, I knew that I had to stop to fill up with gas. (I could never have completed the journey without stopping, as my car only does about 50 miles per gallon on the freeway).

I spent two days driving around Los Angeles, looking at the sights. I must have covered 50 miles! (Los Angeles driving uses up gas at a much faster rate than freeway driving because you have to keep stopping and starting).

On Monday, I returned home by the same route. This time I forgot to check my gas, and I completely ran out on the freeway. After a while, I remembered an old 2-gallon can which I kept in the trunk for emergencies, and this enabled me to get home.



[10]

Scoring Balanced Assessment

Since the approach to scoring used for these tests is unfamiliar to some, we think it worth describing it in outline. It is a widely used approach, which, with scorer training, yields consistent scoring. *The Scoring Guide* is designed to facilitate accurate and reliable scoring, and forms the basis of training scorers.

Below is part of the introductory section, which explains the basis of the design, development, and use of the scoring. It is illustrated with a simple task for Grade 3, *Magazine Cover.*

The essential materials for scoring are provided in the **Resources** Section of the Guide. They comprise, for each task:

- the scoring rubric
- 5 student "training papers," selected to illustrate the use of the rubric across the range of performance. These training papers are provided in both unscored and scored form.
- 10 student "standardizing papers". These papers, too, are provided in both unscored and scored form. They are provided to help scorers practice applying the rubric to achieve consistent scoring.

It should be recognized that no written rubric can cover all possible student responses. Scorers must sometimes use their professional and mathematical judgment in following the rubric. With scoring training using these materials, and appropriate monitoring of scorers, high reliability can be achieved.

Principles of Scoring

Performance assessment provides students with an opportunity *to show what they know, understand, and can do* on a balanced sample of tasks from the range and variety that constitute mathematical performance. Scoring is where we value the different aspects of that performance.

All assessment embodies value judgments. For example, if one values only accuracy in computation and manipulation, credit will only be given for those aspects of performance. If, however, one also values students' abilities to decide how to tackle an unfamiliar problem, or interpret and evaluate solutions, or communicate results and reasoning to others, credit will be given for these aspects as well. In performance assessment, these value judgments are explicit, as set out in the scoring rubrics for each task.

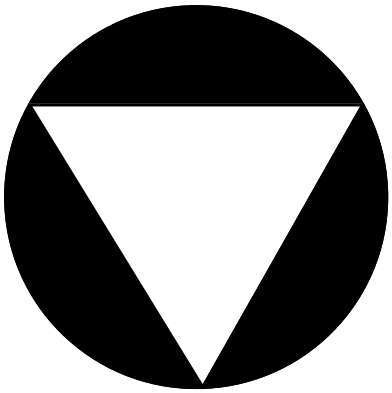
The scoring rubric for each task is developed along with the task. This process involves detailed discussion of the elements of performance that the task demands, as well as analysis of samples of student work. The rubric represents one carefully considered evaluation of student performances on the task. While other approaches are possible, consistency of scoring requires that all scorers apply this rubric carefully and conscientiously, even if they do not agree with some aspects of it.

Magazine Cover exemplifies the tasks and their scoring.

MAGAZINE COVER

This problem gives you the chance to:
• describe a given geometric pattern

This pattern is to appear on the front cover of the school magazine.



You need to call the magazine editor and describe the pattern as clearly as possible in words so that she can draw it.

Write down what you will say on the phone.

The Scoring Rubrics. These rubrics use a point system. Each rubric is an expansion of the description of the *core elements of performance* of the task, and is listed under the title (as in *Magazine Cover,* shown above).

It proves sensible to choose the total points available for each task to be equal to the length of time (in minutes) it takes a typical successful student to complete the task. The total points are then distributed among the different aspects of performance on the task, so that each aspect is given a weight appropriate to its importance.

This process is illustrated in the rubric design for *Magazine Cover,* which shows the various elements that are credited. The maximum score is 6 because this is a 6- minute task. For each of the two simple technical terms, *circle* and *triangle*, 1 point is assigned. Most students get these points. For each of the key geometric insights on the triangle (*equal sides, point down, touching the circle*), 1 point is given, but further technical terms (e.g., equilateral) are not required at grade 3. For the color contrast, seen as a key feature in the context of the task (cover design), 1 point is given, and 1 point is given for some size information. To keep within the total, a maximum of 6 of the 7 points above may be awarded. Later points are designed to represent more difficult challenges, so as to produce a scale that covers the wide spectrum of performance in mathematics.

	Magazine Cover: Grade 3	points	section points
	The core elements of performance required by this task are • describe a given geometric pattern Based on these, credit for specific aspects of performance should be assigned as follows:		
	A **circle.**	1	
	A **triangle.**	1	
	All corners of triangle on (circumference of) **circle.**	1	
	Triangle is equilateral. Accept: All sides are equal/the same.	1	
	Triangle is standing on one corner. Accept: Upside/going down.	1	
	Describes measurements of circle/triangle.	1	
	Describes color: **black/white.**	1	
	Allow 1 point for each feature correctly described up to a maximum of 6 points.		**6**

In more complex tasks, more than one point may be assigned for an important and substantial part of the task, such as for an explanation. The rubric will give guidance on assigning partial credit (e.g., for a correct but incomplete explanation).

Aggregation and Reporting. When all the student papers for a test have been scored, the information has to be interpreted so that it is in a form that is useful and clear. If the scored student papers are returned to teachers and their students, they will get useful, detailed information on individual strengths and weaknesses – as well as growing understanding of this kind of assessment.

For accountability or other external assessment purposes, standardized scores and a well-defined reporting structure may be needed. For these tests, the CTB-MARS Mathematics Board establishes cut scores between the 4 reporting levels, based on a substantial sample of student work from different districts. The main information which guides this standard-setting are the published standards for these tests, the Board's view of students' performance on them, and a range of statistical information. While these national cut scores will be available to all users, some districts may prefer to use their own standard-setting procedures, based on local standards.

Fuller information is contained in the Balanced Assessment Reference Guide, available from MARS.

The challenges. Teachers with no experience in performance assessment will benefit from professional development support. Scoring training is a good basis to build on.