Technical Report # 42

Content-Related Evidence for Validity for Mathematics Tests: Teacher Review

Martha I. Martinez

Leanne Ketterlin-Geller

Gerald Tindal

University of Oregon

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689
http://brt.uoregon.edu

Abstract

Behavioral Research and Teaching (BRT) has developed a series of mathematics tests to assist local school districts in identifying students in grades 1-8 who may be at risk of not meeting year-end mathematics achievement goals. The tests were developed using the state mathematics standards for the relevant grade levels and administered to students in fall, winter and spring. In an effort to continuously improve the tests as well as to examine the validity of their uses, school staff from local districts participated in piloting and reviews of the tests from 2003-2006. The 2005-2006 teacher review documented in this technical report was designed to systematically capture feedback on all test items based on the appropriateness of language, concepts, and graphics, as well as bias in language or graphics. This review provides content-related validity evidence for the uses of the test results as screening tools.

Content-Related Evidence for Validity for Mathematics Tests: Teacher Review

Over the past few years, researchers at Behavioral Research and Teaching (BRT) have collaborated with administrators and instructional staff at three local school districts to develop a series of mathematics tests for grades 1-8. The primary purposes of the tests are two-fold: (1) to assist school staff in identifying students who may be at risk of not meeting the state's mathematics achievement targets by the end of the year, and (2) to help inform instructional practices early on, so that students performing well below their peers could receive additional support. In 2005, BRT staff initiated a systematic teacher review of all the test items. Teachers from two local school districts participated in the review. This technical report documents the process and initial results of the 2005-2006 teacher review of the 1^{st} – 8^{th} grade BRT mathematics tests.

Theoretical Framework

Assessments can take many forms and are used for a variety of purposes. The mathematics tests that are the subject of this report are a type of formative assessment, i.e., they are designed to inform and possibly change instructional practice. Black and Wiliam (1998) define formative assessment as all assessment activities undertaken by teachers and students when the evidence drawn from such activities is used to adapt teaching to meet student needs. Formative assessment, as Black and Wiliam and others (e.g., Pelligrino, Chudowsky, & Glaser, 2001) define it, is often times situated and initiated at the classroom level by the teacher. However, in this paper, we expand on this definition to include all assessment activities with the intent of informing instructional decisions, regardless of who undertook such efforts. The mathematics tests described in this technical report are used by school districts as screening or benchmarking assessments to provide schools and teachers with information about student performance in the area of mathematics calculations and applications. Schools use this

information to help guide decisions such as identifying students in need of additional instructional supports in mathematics. As such, the tests fit the definition of formative assessments. Insofar as benchmarking exams provide useful information to teachers and assist them in making decisions related to student learning, they represent formative assessment in action (Alonzo, Ketterlin-Geller, & Tindal, in press).

Simply collecting assessment data in and of itself will not result in improved outcomes. Webb (1999) and Black & William (1998) suggest that in order to maximize the effectiveness of formative assessments they should be aligned with curriculum, instruction, and state standards. Wade (2001) also argues that assessment data is most useful in positively impacting educational outcomes when it is valued by school staff and systematically collected and analyzed. Based on Black and Wiliam's review of research on whether formative assessment could lead to improved student learning outcomes, they not only conclude that it can, they also argue that there is significant room for improvement of formative assessment practices. They also lend further credence to the use of standardized measures of assessment that are administered across classrooms (such as the BRT mathematics tests) for formative assessment purposes, and suggest that effective learning is impeded because teachers do not often share assessment questions and methods across classrooms.

The BRT mathematics tests were developed to align with the state's mathematics standards (which should theoretically align with curriculum and instructional delivery in Oregon's schools.) BRT has also helped foster teacher and administrative support for the tests by collaborating with school staff in their development and providing assistance with their administration, and training staff on how to interpret the test data. Moreover, because the tests

are administered district-wide, they address at least one problem that Black & Wiliam (1998) suggest impedes student learning: i.e., that teachers often do not share their assessments.

The purpose of the teacher review of the BRT mathematics tests was to examine the content-related evidence for using these tests to inform instructional practices. Content-related evidence for validity helps bolster the assumption that the tests appropriately and adequately measure the subject matter (or academic content and skills) that they purport to measure and in a manner that is consistent with the purposes for which they were constructed. The Test Standards (1999) define validity as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" and content-related evidence as one form or type of evidence based on test content that falls within the larger unitary concept of validity (p. 184).

In 2005-2006, BRT researchers conducted a teacher review to examine the content-related evidence for the use of the BRT mathematics tests. In the teacher review, teachers were asked to examine each test item in terms of four criteria: (a) the appropriateness of the language used, (b) the appropriateness of the concepts tested, (c) the appropriateness of the graphics used to represent the concept being tested, and (d) whether there was bias in the language or graphics used. This report documents the teacher review of the test items.

Methods

In the 2005-06 academic year, BRT conducted a teacher review that was designed to systematically capture teacher feedback on all test items. In the fall, BRT contacted three local school districts that have been using the BRT mathematics tests to request their assistance in conducting a systematic, item-by-item review of all the tests. The review included mathematics

tests from grades 1-8, which consisted of 100-150 test items per grade level. Following is a detailed description of the number of tests and test items that teachers reviewed.

In each of grades 2-8, teachers reviewed a total of six mathematics tests: two for fall, two for winter, and two for spring. The tests were designed to be delivered in pairs each academic term (fall, winter, and spring), with students receiving one test on calculations and the other on applications. Thus, there were a total of three calculations tests and three applications tests for each grade, 2-8. Each test for grades 2-8 consisted of 25 test items, yielding 150 test items per grade level for teacher review. Because grade 1 had only one fall test (calculations), teachers reviewed a total of five first-grade tests (three calculations and two applications tests) and 125 test items for grade 1. (See Table 1 for a summary of the number of mathematics test items reviewed.)

Table 1

Number of Test Items Reviewed Per Grade Level

	Fall	Winter	Spring	Total
Grade 1 Total	25	50	50	125
Calculations	25	25	25	75
Applications	0	25	25	50
Grades 2-8 Total	50	50	50	150
Calculations	25	25	25	75
Applications	25	25	25	75

Note: Figures for Grades 2-8 indicate the number of test items in each grade level, not the cumulative number of test items across grades 2-8.

The teachers were asked to evaluate each test item in terms of the following four criteria: (a) the appropriateness of the language used, (b) the appropriateness of the concepts tested, (c) the appropriateness of the graphics used to represent the concept being tested, and (d) whether there was bias in the language or graphics used. In addition, teachers were instructed to provide their overall evaluation of each test for the specific grade level reviewed in terms of similar criteria.

Setting and Participants

Of the three districts initially contacted, staff at two district offices identified teachers for the review. In December 2005, BRT staff met with the teacher reviewers to discuss the purpose and goals of the review. The meeting was held at one of the district offices. A total of 17 teachers attended the orientation and review meeting, representing sixteen schools across the two districts. The teachers were compensated by their districts for their attendance at the meeting.

Teachers who attended the meeting filled out a "Reviewer Information Form" that asked for background information such as the teacher's gender, ethnicity, grade level taught, and the total years he/she had been teaching (See Appendix A). Most of the teachers identified themselves as white or Caucasian (n=14). Fifteen of the 17 teachers were female, and the average number of years taught for the group was 15 years. See Table 2 for demographic information on the teachers who attended the review meeting.

Table 2

Demographic Information on the Teachers Attending the Review Meeting

	District A	District B	Total
Participants	13	4	17
Gender			
Female	11	4	15
Male	2	0	2
Ethnicity			
White	10	4	14
Chinese	1	0	1
Mixed	1	1	2
Declined to state	1	0	1
Average years teaching	15	13	15

Procedures

The meeting with BRT staff was designed to provide teachers with a context for the review, instructions regarding the review procedures, and time to actually complete the test review. After providing a brief history of the development of the tests, it was explained that the purpose of this review was for teachers to evaluate every test item in terms of four criteria: (a) the appropriateness of the language used, (b) the appropriateness of the concepts tested, (c) the appropriateness of the graphics used to represent the concept being tested, and (d) whether there was bias in the language or graphics used. Appendix B includes the presentation outline that BRT used to frame the meeting and explain the review procedures that teachers were asked to

use. Teachers were instructed to use the review forms (see Appendix C) that were provided to review each test item systematically in terms of the four criteria, as well as to offer any suggestions for how to improve specific items. The review form further explained the questions BRT wished the teachers to consider regarding each of the criteria. For example, following "appropriateness of language", the following questions appeared:

- Are the question and response options written so that students in the assigned grade can understand the meaning of the problem?
- Is the vocabulary written at the appropriate grade level?

In addition, teachers were instructed to provide their overall evaluation of the tests in terms of the same four criteria and the appropriateness of two other factors: the format used and the directions given to students. (See Appendix D for the Overall Ratings form that teachers received and were instructed to use for this purpose.) Each teacher was asked to review all mathematics tests for the grade level he/she was currently teaching. Most teachers completed their reviews on the same day of the meeting with BRT staff. However, some teachers chose to complete their reviews on their own.

Teachers were asked to review the mathematics tests from a grade level at which they had expertise. Most teachers reviewed tests for grades they were currently teaching or one level removed. For example, a 7^{th} grade teacher reviewed tests for both 6^{th} and 7^{th} grades. In addition, two teachers taught students in all elementary grade levels, as either a Title 1 teacher or a school-wide mathematics facilitator. These teachers reviewed tests for 2^{nd} grade, a grade level that was unrepresented among teachers with specific grade level assignments.

BRT received a total of 16 grade level teacher review packets. About 90% (14 out of 16) of the packets were completed and returned within a month of the review meeting. The remaining

two packets were returned one to two months later. The 16 packets came from 14 teachers. Two teachers (a Title 1 teacher and a 7th grade teacher) completed two packets each, and two teachers from the original group who attended the orientation meeting (one 1st grade teacher and one 8th grade teacher) did not return packets. Nevertheless, all grade level tests were reviewed by at least one teacher. Table 3 in the Results section provides a summary of the number of reviewers per grade level.

Data management

All teacher comments were entered into an MS Access™ database in a table called "Teacher Review" in the BRT Math Database. A record was created in the database for each test question that received a teacher comment. The metadata describing the meaning of each database field is provided in Appendix E. In brief, if the teacher categorized his/her comment as having to do with language, concepts, graphics, or bias, the categorization was maintained in the database record. All general comments on the questions were captured in a database field called Suggestions.

After changes were made to the tests, two additional Yes/No database fields were created. One field captured whether BRT made a change to the item, a second field indicated whether the change made responded to the teacher's feedback. Appendix F provides an example of the type of data collected and the manner in which it was captured in the database.

Education graduate students with mathematics teaching experience individually reviewed every comment that was provided by teachers and made recommendations on whether the comments fit one of three categories: comments on *grade-level appropriateness,* comments on *balance of representation,* and *all other comments*. Comments on grade-level appropriateness included concerns that an item might not be taught at the grade level being tested and comments

that an item was too easy or too difficult. Comments on balance of representation included concerns that a mathematical concept might be over- or under- represented on an assessment. All other comments related to clarity of wording, potential bias, and graphics. The lead researcher on the project made the final decision about the categorization of the comments, and whether any changes to the tests were warranted.

Results

In total, BRT received 1,524 responses pertaining to 982 test items from the teacher reviewers. Of this total, teachers noted 726 issues regarding the four categories of language, concepts, graphics and bias that they were asked to consider in the review. The remainder of the feedback came in the form of comments in the "Suggestions" column. Teacher feedback on the test items ranged from flagging an item as problematic in one of the four areas to providing comments about the nature of the problem and suggestions for how to improve the item. For example, on item 30 of the 1st grade spring applications test, one teacher noted a problem concerning language. The comments in the "Language" column read, "Could the girls names be easier to read?" The same teacher then included the following comment in the "Suggestions" column, "More common names. Pat, Sue, Lori, etc." Thus, as this example illustrates, teachers sometimes provided more than one response for a given item, and frequently the comments in the "Suggestions" column clarified or elaborated upon concerns that teachers noted regarding the four review categories for a given item. Table 3 summarizes the number of reviewers, the total feedback received, and the total item-level feedback by the test grade level reviewed.

Table 3

Summary of Total Feedback Received from Teacher Reviewers

		Feedback Received	
Test Grade	Number of Reviewers	Total	Item-level Total
First	2	94	67
Second	2	167	109
Third	3	284	199
Fourth	3	279	179
Fifth	2	171	103
Sixth	1	157	96
Seventh	2	247	128
Eighth	1	186	101
Total	16	1524	982

Of the four categories explicitly under review, bias was the least frequently identified as a concern: only 27 (or less than 4%) of the 726 total responses in these four categorical areas, and less than 3% of the total concerns cited with specific items. The remaining three categories were fairly comparable in terms of the frequency of their appearance in teacher responses. Table 4 provides information about the distribution of the feedback across the four specific areas under review as well as the general "suggestions" category.

Table 4

Frequency of Teacher Feedback by Test Grade and Review Categories

Test Grade	Language	Concepts	Graphics	Bias	Suggestions
First	5	35	18	4	32
Second	30	18	24	5	90
Third	32	9	57	1	185
Fourth	72	30	64	7	106
Fifth	40	27	29	4	71
Sixth	0	0	0	0	96
Seventh	14	53	56	2	122
Eighth	38	38	10	4	96
Total	231	210	258	27	798

Numerous changes were made to the mathematics tests based on the data collected during this teacher review. In the interest of revising the spring tests first (which were due to be administered shortly after the teacher review), preliminary data analyses focused primarily on the feedback pertaining to the spring tests. Revisions to the fall and winter tests occurred after this report was drafted. Thus, this technical report summarizes the revisions to the spring tests only.

Preliminary analyses concentrated on three review areas: graphics, language, and bias. Teacher feedback regarding the appropriateness of the concepts addressed in the mathematics items primarily centered on grade-level appropriateness. In addition, some teachers expressed concerns regarding grade-level appropriateness in the suggestions column. However, a separate BRT mathematics alignment study is examining grade-level appropriateness and balance of

representation; therefore, changes in these categories were postponed until data from both the teacher review and alignment study could be addressed simultaneously.

All other suggested changes were reviewed first by graduate students with expertise in teaching mathematics, then the lead BRT researcher on the project, in order to make appropriate revisions to the test items. After changes were made, BRT administrative staff reviewed all test protocols item-by-item to record if changes had been made in response to teacher comments. For each record (an item commented on by an individual teacher), the staff member marked in the database in a Yes/No field if the item had been changed and whether the change responded to a teacher comment. Finally, each change made to the test item was summarized in a *Comments* field in the database.

Teachers provided feedback on 982 items. Of these total items, 366 (37%) were on the spring tests. BRT made changes to 142 items on the spring tests. Over half of these changed items (76) specifically addressed the feedback teachers provided. The changes made to the spring mathematics tests in the areas of graphics, language and bias are summarized below.

In the area of graphics:

- Mathematics items were reformatted so that each question and its response choice appear on the same page.
- Line spacing was increased: both between mathematics items and between response choices.
- Graphics were made more legible.
- The font style was changed to san-serif (Tahoma) to increase reliability.
- Response choice formats were revised to reflect consistency across the test and to simplify navigation.

In the area of language:

- Questions written as incomplete sentences were reworded as direct questions, whenever possible.
- Linguistic complexity was reduced when such complexity was unnecessary for the particular item.
- The use of language/terminology across similar items was made more consistent.

In the area of bias less familiar terminology was replaced with more common words or phrases whenever possible.

Discussion

Although BRT has solicited teacher feedback on the mathematics tests in previous years, this was the first time that we had conducted a comprehensive and systematic review of each test item. As a result of this process, BRT was able to capture feedback from teachers on four specific areas for each test item: the appropriateness of the language, the concepts, and the graphics, as well as bias in language or graphics. In the interest of revising the test before the next administration in spring, data analysis focused on the spring tests. Subsequent analysis of the teacher review data in tandem with an analysis of the data collected from the alignment study will yield an even more complete picture of the strength of the validity evidence for the uses of the tests and help allow BRT to continue to refine the tests to maximize their instructional effectiveness.

We are increasing relying on assessment systems for a variety of purposes: accountability, instructional guidance, progress monitoring, and evaluation. Thus, the need to show evidence for the validity of these uses is imperative. The purpose of this teacher review was to examine the content-related validity evidence that support the uses of the BRT

mathematics tests to help teachers identify students at-risk of not meeting year-end mathematics achievement goals. Although the validity evidence collected thus far is by no means conclusive, it is nevertheless an important step in the test development process.

References

Alonzo, J., Tindal, G., & Ketterlin-Geller, L.R. (in press). General outcome measures of basic skills in reading and math. In L. Florian (Ed.), *Handbook of Special Education*. Thousand Oaks, CA: Sage.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D. C.: American Psychological Association.

Black, P. & Wiliam, D. (1998, October). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan.*

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). The nature of assessment and reasoning. In *Knowing what students know: The science and design of educational assess*ment (pp. 37-54). Washington, DC: National Academy Press.

Wade, H. (2001). *Data inquiry and analysis for educational reform. ERIC Digest.* Eugene, OR: ERIC Clearinghouse on Educational Management. (ERIC Document Reproduction Service No. ED461911)

Webb, N.L. (1999). *Alignment of science and mathematics standards and assessments in four states*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 18. Madison: University of Wisconsin – Madison.

Appendix A

Math Test Review
Reviewer Information Form

Please complete the following by December 19, 2005. Send or fax (541-346-5689) to:

Behavioral Research and Teaching
232 College of Education
5262 University of Oregon
Eugene, OR 97403-5262
Phone: (541) 346-0071; Fax: (541) 346-5689

Name_____
Email address
School
School Address
School Phone Number
School FAX Number
Current Grade Assignment
Years Experience Teaching this Grade
Previous Grade Assignments
Total Years Teaching
Years to Retirement
Gender
Ethnicity
Degree(s)/Certification(s) (and year)

Math Test Review

Collaboration with Eugene 4J and Bethel School Districts and Behavioral Research and Teaching

Agenda

- Review the purpose of the test
- Discuss the purpose of the test review
- Describe the procedures for the review
- Review the tests

Purpose of the Test

- Make decisions about student math proficiency
- Test scores can be used to:
  - Identify students who may need additional services in mathematics
  - Monitor progress toward reaching benchmarks

Purpose of the Test Review

- Develop an appropriate testing system that will help you make instructional and systems-level decisions
  - Identify strengths and weaknesses in the current tests for making decisions
  - Provide specific feedback for improvements at the item level

Current Work on the Tests

- Alignment with the state standards
- Test reviews by teachers

Previous Teacher Review

What it Entailed

Common Themes

Unexpected Outcomes

Current BRT Review

- Item-to-Standard Review
- Process
  - Independent analyses (done)
  - Reconcile differences (next step)

Current Teacher Review

- How it differs: systematic item review
- Overlap?
- Feedback considered in broader context
  - Previous teacher review
  - BRT item-standard review
  - Current teacher review
  - Purpose of the test/items

Purpose of Reviews

- Develop an **appropriate** testing system that will help you make instructional and systems-level decisions
- Perfect Test vs. More Appropriate Test
- Documentation of development

Analyses of the Results

- Results of the test review will be implemented for spring administration
- Results will be reported in technical report
- Administration manual will be drafted for spring to address additional issues that surface

Future Plans for the Math Test

- Reporting system
- Format options:
  - Paper-pencil format
  - Computer-based administration
  - Computer-adapted administration

Procedures for the Review

- Components in the review:
  - Appropriateness of language
  - Appropriateness of concepts
  - Appropriateness of graphics
  - Item bias

Appropriateness of Language

The seating arrangement for a choral concert resembles a pyramid with one student on the top bleacher two on the second and so on. If there are 28 students in the arrangement, how many rows are there?

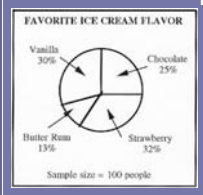
A. 6 B. 7 C. 8 D. 14

Appropriateness of Concepts

Arrange the digits 4,3,7,0, and 6 to form the number with the GREATEST value.

A. 74,630
B. 76,340
C. 76,403
D. 76,430

Appropriateness of Graphics



How many people selected Strawberry as their favorite ice cream flavor?

A. 30
B. 60
C. 70
D. 120

Item Bias

The most reasonable unit to measure the length of a skateboard is the

A. Millimeter
B. Decimeter
C. Centimeter
D. Kilometer

Putting the Pieces Together

- Review the items
- Complete the item review form
- Discuss the items with the members at your table

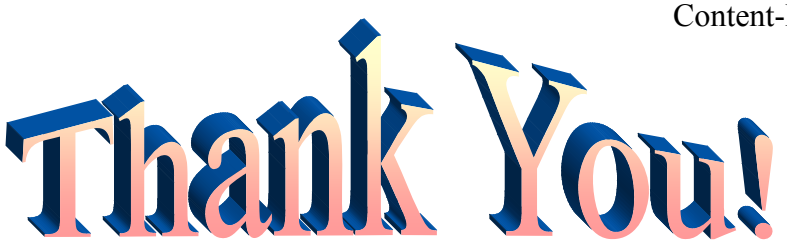
What are your thoughts?

- Questions or comments?
- Plan for the review…

Math Test Review

Collaboration with Eugene 4J and
Bethel School Districts and
Behavioral Research and Teaching

Appendix C



Thank you for reviewing math items for the district assessment. We appreciate your time and effort to make this test the most appropriate for your students. We would like to receive your feedback by Monday, December 19, 2005.

As you review each item, please consider the following issues. Any suggestions would also be appreciated.

- **Appropriateness of language**: Are the question and response options written so that students in the assigned grade can understand the meaning of the problem? Is the vocabulary written at the appropriate grade level?

- **Appropriateness of concepts**: Can students in the assigned grade complete the task? Is this information taught within the normal curriculum of the grade?

- **Appropriateness of graphics**: Will the students be confused by any of the graphics included in the item? Do the graphics appropriately represent the concept being tested? Are the graphics distracting? Is the graphic clear?

- **Bias in language or graphics**: Does the item require background knowledge unrelated to the concept being tested that would differ for students with different backgrounds? Is the language sensitive to students from diverse backgrounds?

You may use the chart provided as you review the items. If you would prefer, you can make your comments or suggestions directly on the test.

Please return your comments and the tests to:

Behavioral Research and Teaching
230 Education
5262 University of Oregon
Eugene, OR 97403-5262

*Test Code: _____

Grade Level: _____

Question Number	**Language?**	**Concepts?**	Graphics?	**Bias?**	**Suggestions**
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					

*This is the first page of a multi-page chart that included rows for 50 test items. Reviewers received one form for each test they reviewed. Tests were coded to indicate the term (fall, winter or spring), grade level, and type of test (computations or applications).

Appendix D

Overall Ratings of Difficulty

Test Code: _____
Grade Level: _____

Please consider the overall test when responding to the question below. Circle the choice that most closely aligns with your impressions of the test.

1. How appropriate is the *language* used for students in the assigned grade?

not at all appropriate	*somewhat appropriate*	*appropriate*	*extremely appropriate*

2. How appropriate is the *format* of the math items for students in the assigned grade?

not at all appropriate	*somewhat appropriate*	*appropriate*	*extremely appropriate*

3. How appropriate are the *concepts* of the math items for students in the assigned grade?

not at all appropriate	*somewhat appropriate*	*appropriate*	*extremely appropriate*

4. How appropriate are the *graphics* used in the items for students in the assigned grade?

not at all appropriate	*somewhat appropriate*	*appropriate*	*extremely appropriate*

5. How clear are the *directions* for students in the assigned grade?

not at all clear	*somewhat clear*	*clear*	*extremely clear*

6. How *biased* are the items based on the experiences of the students in the assigned grade?

extremely biased	*somewhat biased*	*not at all biased*

7. What suggestions do you have to improve the test? Please use the back of this form if you need additional space to provide your answer.

Appendix E. Metadata for table Teacher Review in BRT Math Database.

Field Name	Data Type	Field Description
ItemID	AutoNumber	A unique code that identifies the comments that a single reviewer made on a single test item
TestGrade	Number	This is the grade level written on the test at the top of the page. The tests are grades 1-8, so only the numbers 1 to 8 should be entered in this field.
SeasonID	Number	This is the season (fall, winter, spring) of the test, which is written at the top of the test page.
ItemNumber	Number	This field is the test item (or question) number. For example, the first grade test has 20 questions. Each question or item is given a separate line in the database.
ReviewerID	Number	This is a unique code that identifies the teacher that provided each comment.
Language	Memo	Any comments written in the Language column from the feedback paper provided by BRT to teachers. These should be comments on the wording of assessment questions.
Concepts	Memo	Any comments written in the Concepts column from the feedback paper provided by BRT to teachers. These should be comments on the mathematical concepts being assessed.
Graphics	Memo	Any comments written in the Graphics column from the feedback paper provided by BRT to teachers. These should be comments on the clarity, design, and appropriateness of figures, tables, and other graphics.
Bias	Memo	Any comments written in the Bias column from the feedback paper provided by BRT to teachers. These should be comments on any potential bias inherent in assessment questions.
Suggestions	Memo	Any comments written in the Suggestions column from the feedback paper provided by BRT to teachers. Also, ANY comments written on test papers and not categorized by a teacher as Language, Concepts, Graphics, or Bias

Field Name	Data Type	Field Description
Markings on Item?	Yes/No	YES -- if the teacher marked the item only by highlighting or circling, but didn't provide any written comments. NO -- If there are markings *in addition to written comments* or no markings at all.
QuestionPaper?	Yes/No	YES -- If the comment is written on a question paper (actual test) NO -- other
FeedbackPaper?	Yes/No	YES -- if the comments are from the feedback packet provided by BRT. NO -- other
BRT Made Changes to Item?	Yes/No	The item (question and/or answers) has been changed.
Change responded to reviewer comment?	Yes/No	The change in the item is directly related to a teacher comment.
Summary of Change	Memo	Summary of the change made to the item.

Appendix F. Reviewer comments and changes made to first grade spring assessments.

Item Number	Reviewer ID	Concepts	Graphics	Bias	Suggestions	BRT made change to Item?	Change responded to reviewer comment?	Summary of Change
21	6				ok.	TRUE	FALSE	Question changed from "Fill in the blank" to "Which number is missing." Font bigger, spaced evenly, formatted with the number and question in horizontal row.
21	6				21 is crossed out.	TRUE	FALSE	Question changed from "Fill in the blank" to "Which number is missing." Font bigger, spaced evenly, formatted with the number and question in horizontal row.
21	12	ok	Too close too other problems and print should be larger.			TRUE	TRUE	Question changed from "Fill in the blank" to "Which number is missing." Font bigger, spaced evenly, formatted with the number and question in horizontal row.
22	6				ok.	TRUE	FALSE	Question changed from "Fill in the blank" to "Which number is missing." Font bigger, spaced evenly, formatted with the number and question in horizontal row.
22	12	ok	need more		Why not start with 5 since you want to see if they can count by 5.	TRUE	FALSE	Question changed from "Fill in the blank" to "Which number is missing." Font bigger, spaced evenly, formatted with the number and question in horizontal row.
23	6				ok.	TRUE	FALSE	Question changed from "Fill in the blank" to "Which number is missing." Font bigger, spaced evenly, formatted with the number and question in horizontal row.

Item Number	Reviewer ID	Concepts	Graphics	Bias	Suggestions	BRT made change to Item?	Change responded to reviewer comment?	Summary of Change
23	12	ok	space between problems			TRUE	FALSE	Question changed from "Fill in the blank" to "Which number is missing." Font bigger, spaced evenly, formatted with the number and question in horizontal row.
24	6				How do you want answers marked? Circled? Ok. See test page.	TRUE	TRUE	Question changed from "Add the two coins and mark your answer" to "Add the two coins. Circle your answer." Formatting clearer, problem number and question in horizontal row, then coins, then answer selections.
24	12	ok	clean up unclear coins			TRUE	TRUE	Question changed from "Add the two coins and mark your answer" to "Add the two coins. Circle your answer." Formatting clearer, problem number and question in horizontal row, then coins, then answer selections.
25	6				How do you want answers marked? Circled? Ok. See test page.	TRUE	FALSE	Answer selections in vertical row instead of horizontal, much clearer.
25	12	ok with clearer graphic	Circle and #'s need to be bigger.		Replace stop with land as that's what is said more often. Since this probability using a graphic like drawn in exampe - might more clearly let one know if child understands probability and not INCREASE SIZE. Good that arrow is at a neutral spot. See	TRUE	FALSE	Answer selections in vertical row instead of horizontal, much clearer.

Item Number	Reviewer ID	Concepts	Graphics	Bias	Suggestions	BRT made change to Item?	Change responded to reviewer comment?	Summary of Change
					paper.			
26	5				ok. How do you want answers marked? Circled? Ok. See test page.	TRUE	TRUE	Question changed from "Which shape is the triangle?" to "Circle the triangle below."
26	12	ok if not testing the reading of the word.		Needs to be clearer.		TRUE	FALSE	Question changed from "Which shape is the triangle?" to "Circle the triangle below."
27	12	ok	ok	ok		TRUE	FALSE	Question changed from "Which line is longer?" to "Circle the longer line." Formatting clearer with number and question in horizontal row.
27	6				ok. How do you want answers marked? Circled? See test page.	TRUE	TRUE	Question changed from "Which line is longer?" to "Circle the longer line." Formatting clearer with number and question in horizontal row.
28	12	ok	Lines inside shapes are distracting.	ok		TRUE	FALSE	Answer options listed vertically, much clearer.
28	6				ok.	TRUE	FALSE	Answer options listed vertically, much clearer.

Item Number	Reviewer ID	Concepts	Graphics	Bias	Suggestions	BRT made change to Item?	Change responded to reviewer comment?	Summary of Change
29	12	ok	Larger calendar squares!		How is the demonstration question a demonstration for the questions on this page?	TRUE	FALSE	Calendar much clearer, equal spacing. The word October added into question for clarity. Formatting much clearer, question number before question, answer options vertical.
29	6		Make chart boxes equal.		Couldn't the calendar grid look better - equal space for each row?	TRUE	TRUE	Calendar much clearer, equal spacing. The word October added into question for clarity. Formatting much clearer, question number before question, answer options vertical.
30	12	Wow? NO. Too many skills in one problem.			If this is to measure a child's chart interpretation then a bar graphic is more age appropriate. Too many skills for one question. 1. Chart reading?? Is this the skill tested? 2 subtraction 3 double digit subtraction 4 douible digit subtraction with borrowing 55-39 5 take the bottom # put on top then subtract 6 reading 2003 2004	TRUE	FALSE	Names are more common, font larger.
30	6				More common names. Pat, Sue, Lori, etc.	TRUE	TRUE	Names are more common, font larger.