Recommendations for Building a Valid Benchmark Assessment System: Interim Report to the Jackson Public Schools

CRESST Report 723

David Niemi, Julia Vallone, Jia Wang, and Noelle Griffin
CRESST/University of California, Los Angeles

July, 2007

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Stupski Foundation's Support of Assessment Development: Jackson Public Schools program, Award # 20050546. The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the Stupski Foundation.

RECOMMENDATIONS FOR BUILDING A VALID BENCHMARK ASSESSMENT SYSTEM: INTERIM REPORT TO THE JACKSON PUBLIC SCHOOLS

David Niemi, Julia Vallone, Jia Wang, and Noelle Griffin

CRESST/University of California, Los Angeles

Abstract

Many districts and schools across the U. S. have begun to develop and administer assessments to complement state testing systems and provide additional information to monitor curriculum, instruction and schools. In advance of this trend, the Jackson Public Schools (JPS) district has had a district benchmark testing system in place for many years. To complement and enhance the capabilities of district and school staff, the Stupski Foundation and CRESST (National Center for Research on Evaluation, Standards, and Student Testing at UCLA) worked out an agreement for CRESST to provide expert review and recommendations to improve the technical quality of the district's benchmark tests. This report (which represents the first of two deliverables on this project) focuses on assessment development and is consistent with the district goal of increasing the predictive ability of the assessments for students' state test performance, as well as secondary goals.

Many districts and schools across the U. S. have begun to develop and administer assessments to complement state testing systems and provide additional information to monitor curriculum, instruction and schools. In advance of this trend, the Jackson Public Schools (JPS) district has had a district benchmark testing system in place for many years. The district has successfully implemented mechanisms for developing, distributing, administering, scoring, and reporting on benchmark tests in multiple subject areas several times per year. This is a considerable achievement, requiring significant expertise on the part of district and school staff.

To complement and enhance the capabilities of district and school staff, the Stupski Foundation and CRESST (National Center for Research on Evaluation, Standards, and Student Testing at UCLA) worked out an agreement for CRESST to provide expert review and recommendations to improve the technical quality of the

district's benchmark tests. In response to requests by the district to provide a deliverable to JPS earlier than the original reporting date (June 30, 2006), the scope and sequence of CRESST's work was revised to include two deliverables, of which the contents of this report represent the first. This report focuses on assessment development and is consistent with the district goal of increasing the predictive ability of the assessments for students' state test performance, as well as secondary goals.

The recommendations in this report cover the Grades 3-12 interim assessment development and validation process in elementary Language Arts, elementary Mathematics, Algebra I, English I, Biology I, and U.S. History. This work is intended to ensure that both process and content are consistent with widely-accepted assessment development standards (e.g., AERA, APA, & NCME, 1999; Hambleton, 1996; Linn, 1993).

To meet these standards, the following conditions must be achieved:

1. The purposes of the assessment are clearly defined.
2. The domain to be assessed is clearly specified.
3. Item development and selection procedures as well as administration and scoring procedures are accurately documented.
4. Alignment: there is credible evidence on the match between assessments and domain specifications.
5. Procedures for analyzing, using, and reporting data meet appropriate technical standards.
6. Validation: evidence is assembled to support the intended interpretations and uses of the assessments.

Our recommendations will therefore describe a) how these conditions apply to the JPS benchmark tests, and b) what must be done to insure that the conditions are met; that is, to insure that the tests are of sufficiently high quality to meet basic professional and educational standards, as well as serving the district's purposes. We assume that CRESST and Stupski will continue to work with the district to assist in implementing these recommendations.

Assessment Purposes

High quality assessment is not possible without clearly identifying the purposes of the assessments. Assessments have to be judged against their intended uses. There is no absolute criterion that can be used to judge assessments and assessment systems that do not have clearly spelled out purposes. It is not possible to say, for example, that a given test is good for any and all intents and purposes; it is only possible to say, based on

evidence, what purposes the test is valid for. Furthermore, professional assessment standards require that assessments be validated for all their intended uses (e.g., AERA, APA, NCME; CRESST; National Research Council, 2001). A clear statement of assessment purposes also provides essential guidance for test and assessment item developers. Different purposes may require different content coverage, different types of items, etc. Thus it is critical to identify with as much precision as possible how assessment information is to be used, and to validate the assessments for those uses.

According to the district (based on correspondence from Dr. Potter dated 4-4-06), the prioritized uses of the benchmark test data are as follows:

a. Predictive ability of the assessments (how well would a student perform on the State Assessment in that area)

b. School evaluation: How well are students in a school being prepared to be successful on the state assessments?

c. Student Performance: Data will continue to be used as a key component of student grades.

d. Overall Accountability of System: Are students in Jackson mastering the State assessments?

These uses suggest that some additional analysis by the district is needed. Predictive ability of the assessment is not a use per se; rather, it's a quality of the assessments. For example, college admissions tests are supposed to predict future performance in college, but the purpose of these tests is to make decisions about who should be admitted to college. Part of the validity evidence for these tests consists of data on whether students who perform well on the test also do well in college. Similar correlational evidence should be obtained for the JPS benchmark tests (i.e., do scores on the benchmark tests correlate highly with state test scores?), but the uses of the benchmark test scores still need to be spelled out. For example, if the assessments prove to be highly predictive, how will that information be used, and by whom? These questions need to be answered to address some of the design and validity questions discussed below.

School evaluation, student grading, and system accountability are mentioned as secondary uses, but these may turn out to be the main uses of the benchmark test data. Presumably, state test scores can also be used to address these purposes. If the tests are supposed to predict performance on the state test, there is no compelling reason to

administer benchmarks after or just before the state test. Testing after the state test renders prediction meaningless. Testing just before the state test raises difficult questions: What is the value of giving a predictive test shortly before the state test? Why not just wait to see how students do on the state test?

With respect to school evaluation, grading, and system accountability as purposes, there are many questions still to be answered, for example, How will scores be used to evaluate schools or the system? Will the percentage of proficient students be the primary measure? Or improvements in those percentages over time? How will the prior achievement levels of students and other variables influencing performance be accounted for in evaluating schools? If the tests are used for grading purposes, will all students have an equal opportunity to learn the content? Will the test cover a representative sample of the curriculum presented to students? These questions are probably best addressed through ongoing consultations between CRESST and the district.

If the benchmark tests are explicitly designed to correlate highly with the state tests, this will mean that the district is committed to using a state-test-like measure to evaluate schools, student performance and overall system accountability. The question then will be whether a measure based on the state test is valid for these purposes. Will there be other assessments to account for the fact that a multiple-choice state test designed for accountability purposes cannot adequately address all important types of mathematics learning? How will the district insure that curricula and instruction are not unnecessarily narrowed to the content that is assessed on the state test?

Domain Specification

Models of learning, cognition, and expert performance in the domain should be used to develop assessment specifications. This is one of the strongest and most consistent recommendations of measurement and assessment by experts and professional groups. For example, the AERA/APA/NCME's *Standards for Educational and Psychological Testing* (1999) state that assessment development should have a scientific basis; e.g., specification of the domain to be assessed should draw on cognitive research in the domain, and task design should be informed by cognitive analysis and empirical testing of tasks.

To insure that assessments reflect the content domain to be assessed, assessment design and validation should be preceded by a cognitive analysis of the domain or by an analysis of the cognitive demands of the performances to be assessed. To analyze the "cognitive demands" of a performance means to specify the knowledge and mental skills required to complete that performance successfully. A recent report by CRESST (Herman & Baker, 2005) suggested that analysis of the domain should at least take into account, in addition to the specific content of individual state standards, both the intellectual demands of the domain and the "big ideas" or key principles underlying the content standards. An analysis of "key principles" obviously pre-supposes a theoretical understanding of the content area under consideration (i.e., language arts, math, social studies). Cognitive demands can be viewed through a number of models/categorization systems. In the above referenced article, Baker and Herman propose a framework based on the work of Webb (1997), which identifies four general categories of cognitive demands: recall, conceptual understanding, problem solving/schematic thinking, and strategic thinking/transfer. Although the primary goal of the JPS benchmark tests is predicting state test performance, for the tests to have any additional valid uses by teachers, schools, or administrators these "big ideas" need to be considered in test design and validation.

If the district tests are based on the state test, the district must assume that the state or its contractors has conducted the requisite analyses. This is a questionable assumption at best, but one that the district will have to make, given its desire to mimic the state testing system. The district's domain specifications will have to come from the state's test blueprints. The district should still make every effort to validate the uses of its own assessments, within the context of the big ideas addressed in the assessments.

Given the primary focus on the assessments as a predictor of state tests, to enhance the degree to which the benchmark tests correlate with the state test, the benchmark tests should adhere as closely as possible to the state test blueprints with respect to content, number of items per topic, and cognitive demands and format of the items. However, as noted above, if other uses are considered for the test the district should consider making some adjustments to this content coverage. It is assumed that the testing schedule will remain as it is, with tests every nine weeks. Each test should cover major state test content that is also covered by district curricula within the nine-week time frame. If the main purpose of the test is not to provide comprehensive diagnostic information, the test does not need to be overly long. It is estimated that 30 items should provide a reasonable sample of state test content taught within a nine-week period.

It is also critical that the benchmark tests reflect content taught in each nine week period. Students should not be tested on content they have not been taught (except in the case of tests that are intentionally used to measure the transfer of knowledge).

We examined a sample of documents made available to us by JPS. These included Mississippi Curriculum Test (MCT) blueprints along with the benchmark test blueprints and item analyses. The blueprints for the MCT (math) indicate five reporting categories: patterns, algebraic thinking; data analysis, prediction; measurement; geometric concepts; and number sense. The MCT blueprint also shows the numbers of items within each reporting category (all multiple choice). One question we have is how the items that appear on the benchmark tests were determined. The JPS test blueprint for the benchmark test shows the same reporting categories and also indicates the numbers of items for each category.

As mentioned above, given the stated primary goal for the test, the benchmark tests should match as closely as possible the state test blueprints with respect to content, number of items per topic, and cognitive demands and format of the items. In order to provide an example of how to address these issues for each benchmark test, we carried out an analysis of second-grade math (all but cognitive demands could be evaluated). We looked at the JPS Test Blueprint and Item Analysis documents for each of the four benchmark tests. Appendix A shows an example of a close analysis of the second-grade math documents. This analysis shows the following:

The number of items per each reporting category found on the benchmark test.

The percentage of items in each reporting category across all fours tests.

The percentage of items from each reporting category assessed on the MCT.

Following our analysis, some issues to consider are:

First of all, when designing each assessment, make sure all objectives are covered somewhere in the 9 benchmark tests. In the Grade 2 math example there are a couple of benchmarks that are not assessed in the benchmark tests.

Also, the percentage of items on the benchmark test ought to reflect the percentages on the MCT. This is not the case for the grade two example given here. Indeed, as you can see in Appendix A, 22% of the items on the benchmark tests come from the data analysis and prediction reporting category, while on the MCT only 15% of the test focuses on these concepts.

Another point about the MCT is that it contains objectives that are not all equally important, and this should be taken into account in benchmark test design. For example, benchmark 36 states: "makes change up to $1.00" and benchmark 13 states: "identifies vocabulary terms for time (e.g., before, after, until)". As a comparison, benchmark 24 states: "use the inverse relationship of addition and subtraction". We would consider the latter benchmark as a more important focus than say, number 13 and would therefore recommend more items on the benchmark test address this idea.

Conducting an analysis of the domain that takes into account both the intellectual demands of the domain and the "big ideas" or key principles underlying the content standards would allow one to make "educated" judgments about what content should be focused on. If the MCT blueprint was clearly focused around a hierarchy of big ideas, it would make selecting the objectives for the benchmark tests easier and more meaningful.

As all objectives are not created equally it would make the most sense to focus on what makes the greatest difference in student performance and long-term learning. To achieve this goal it may be necessary to cut objectives from the benchmark tests and focus more heavily on the concepts that will have the greatest impact once they are mastered. For example, a student who has mastered the concepts of the inverse relationship between addition and subtraction will be more prepared to master subsequent concepts. Other objectives may take too long to teach and the value of a student mastering some content may not be as great as other ideas.

Because state tests such as the MCT often focus on easy to assess concepts, the local-based benchmark tests may want to focus assessing deeper understanding of some of the more important concepts. The MCT Blueprint is designed for one purpose and the benchmark assessment blueprint may have more than one specific application. Thus, perhaps the benchmark assessments should focus on other important concepts and not necessarily only the ones contained in the MCT.

Item and Test Development

If the goal is purely to maximize correlations with the state test, the JPS benchmark tests should be constructed by determining which content targeted by the state test is covered in each instructional period. However, as discussed above the district should consider expanding its blueprints to incorporate important conceptual content not

included in the state test. This can be done without seriously affecting correlations with the state test.

Documenting the Development Process

The district has produced blueprints showing topics and number of items per topic. The list of topics appears to match those in the state test blueprint, but the distribution of items across topics does not, as discussed earlier. Some mismatches between blueprints may be appropriate, but the rationale for each decision should be documented.

In addition to information specified in the blueprint, the district should document its decisions in each of the following categories:

Assessment purposes, domain specifications, and rationale for these specifications

- See Assessment Purposes and Domain Specification sections above.

- Domain specifications should cover:
  1. Intended cognitive demands of tasks (e.g., recall of facts or memorized procedures; complex problem solving; recognition, application, and explanation of concepts; use of concepts to justify problem solving; transfer of knowledge)
  2. Mapping of item and response formats (e.g., selected or constructed response) to targeted knowledge and skills. Number and types of items per sub-domain should be specified.

Item and test development procedures

- Test length

  30 items per test should be sufficient for the stated purposes of the test. These items should focus on the most important knowledge and skills taught before the test. If the district decides that the tests should provide diagnostic information, there would need to be 3-5 items per topic reported.

  Currently a small number of staff are developing a large number of items on a very short time line, which is not likely to result in the highest possible quality of items, despite the evident competence of the staff doing the work. Our recommendation on test length is supported by preliminary analysis of one of the tests (Algebra I), which indicates that internal reliability of the test is reasonably high and that many items can be removed without seriously impairing that reliability.

- Test scheduling

  Currently the benchmark tests are administered every nine weeks. This schedule is not problematic, except for problems associated with administering a predictive test after or just before the state test (discussed earlier). Also, we question whether it is necessary to administer five tests (including the state test) per year to monitor school performance and predict state test performance. Two tests before the state test might provide sufficient information.

- Item development

  Items should be designed as specified in the blueprint by item writers who are knowledgeable about the content and who have been trained to write the appropriate kinds of items. Many guides for writing items are available (e.g., Osterlind); as an example we are appending to this document CRESST's multiple choice item writing guidelines (Appendix B), intended originally for teachers.

Administration procedures and directions should describe the qualifications of administrators, administration procedures, permissible variations in procedures, directions for administrators and test takers, and time limits.

Information about the test

- Students, teachers, and schools should receive advance information about the content and format of the test. For this purpose it is reasonable to release a representative sample of items for each test, but not all the items from every test. Releasing all items from every test is an unusual practice and puts a huge item development burden on district staff. Eventually it will become extremely difficult to create new items that are comparable to but different from all the previously used items. A wiser procedure would be to build up a database of items from which new tests can be drawn each year; this database can be developed over time by not releasing all items from every test. There should also be some common items on tests administered in different years, which will make it possible to compare performance from year to year. (Such comparisons are not possible under the current system.) Finally, if all items are released every year, it is not possible to use item data to improve future tests. At some point in the future, when a very large database of items (2000 or so) have been developed for a course, then a higher percentage of items can be released, if desired.

Scoring procedures

- Documentation on scoring should describe: scoring and training methods for open ended items, possible differential weighting of items, instructions and possibly training for interpreting and using scores, procedures for determining proficiency levels, criterion and/or norming procedures, and scaling procedures, if any. There should be a rubric and examples of performance at different score levels for constructed response tasks.

- In addition, the meaning of scores should be clearly stated, and justification for score interpretations provided.

Student characteristics

- Who is the test intended for and who is it not appropriate for should be outlined. For example, are there assumptions about language ability, prior knowledge, format familiarity, special education students, etc.

Alignment

Alignment generally describes the extent to which tests reflect state standards and assessments. Given JPS's focus on the predictive value of the benchmark assessments for state testing, alignment is a critical component of the assessment system. As described in the recommendations above, issues of alignment are reflected in all stages of assessment development and application, from item design to validation to use of assessment data. An important consideration in designing assessments is thus the integration of some formal evaluation of alignment into the assessment development process. A number of detailed models have been proposed for conducting formal alignment analyses of standards and assessments (Porter & Smithson, 2002; Rothman, Slattery, Vranek & Resnick, 2002; Webb, 1997). These models all share a focus on both actual standards content (i.e., instructional topics covered) and some formulation of cognitive complexity/level of knowledge. That is to say, merely checking whether topic assessment and standards topics line up is necessary but not sufficient, and some additional analysis of item complexity is warranted.

Any of these existing alignment analysis models could be implemented by JPS as part of an alignment-check procedure for the benchmark assessments. However, these formal models can also require significant time/resource demands for full implementation. For example, they can include detailed scoring rubrics that would require additional training, scoring and analysis activities on the part of district staff, a need for numerous district raters, and/or additional consultation costs.

Based on guidelines for best practices in assessment alignment that CRESST has developed (Herman & Baker, 2005), it is recommended that a post-item design alignment review be built into the benchmark assessment process (i.e., after the blueprint and individual assessment items have been designed and selected for inclusion in the benchmark assessments). In lieu of using extensive alignment-analysis models, JPS could at minimum implement a less-elaborated alignment procedure drawing on the body of alignment research. This streamlined process will require the designation of senior reviewers within the district to conduct alignment reviews. The reviewers should have content expertise in the domain under consideration for each assessment, and also will need some specific professional development for consistency in the review process. Inter-rater reliability checks, to ensure the reviewers are using and applying their ratings consistently, should also be a part of the training and reviewing process. The reviewers should not be the same individuals who designed the items, and optimally there would be at least two reviewers per content area.

As part of the alignment review process, the senior reviewers should analyze and identify the match of each item to the proposed standards to be tested in terms of the number of dimensions that incorporate both curricular content and skill level. Either a more formalized rubric could be used, such as those developed for the detailed models cited above, or a more simplistic "check box" approach akin to the domain specification example described earlier in this document (and provided in Appendix A). Whichever specific approach is implemented, it should at minimum incorporate the following dimensions:

> Content checking (i.e., is the item at face value an example of the "what", 'how" and "why" of specific standard(s) ostensibly being tested, as detailed in the blueprint)
>
> Cognitive skills/demands (i.e., what cognitive abilities does the item draw on). As described above, cognitive demands can be conceptualized using a wide number of categorization systems. One formulation CRESST (CITE) has employed, based on the work of Webb (1997) includes four categories: recall/memorization, conceptual understanding, problem solving, and strategic thinking (i.e., transfer of knowledge).
>
> Content coverage (or, specifically, the extent of match between the item and the "key principles" or big ideas underlying the specific standard). This dimension should address whether the item under consideration is a good representation of the core concept underlying the standard, or only peripherally/ superficially linked.

Although there is potential to add additional dimensions into this analysis, such as item difficulty, given the scope of JPS benchmark assessments and the resources needed for review it would be advisable to keep the number of dimensions small.

Optimally, this review process will both provide additional support for test alignment and provide an additional check for identifying discrepant or "poor match" items prior to empirical testing of the items.

Procedures for Analyzing, Reporting, and Using Data

Recommendations in this category were delivered in the June 30, 2006 deliverable (for further information please see CRESST Report 724, "Recommendations for Building a Valid Benchmark Assessment System: Second Report to the Jackson Public Schools." Once items were administered, data could be analyzed to determine which items provide the highest quality information and which items need to be modified or dropped. Appendix C contains an initial analysis of one of the benchmark tests. This is an example of one type of item analysis needed to inform future item and test construction. As the report suggests, the items seem to be measuring one underlying construct, rather than multiple sub-constructs. Also, the difficulty of some items may not be appropriate. The June 30 deliverable provides further information and guidance on how item data can be used to improve the benchmark tests over time. The deliverable also covers what needs to be done to insure that results of the test are correctly interpreted, reported and used.

Validation

Validation is another major focus of the June 30, 2006 deliverable, which covers the following considerations for the benchmark assessments (Herman & Baker, 2005):

> Alignment: Do the assessments reflect state standards and assessments, as well as local curricula and instruction?
>
> Fairness: Do the tests provide an accurate assessment of various subgroups, so that all students' needs can be addressed?
>
> Technical quality: Do the tests provide accurate and reliable information about student performance?
>
> Utility: Do teachers, students and others understand the test results and can they use them to improve teaching and learning?

Cost/feasibility: Is the assessment system cost effective and efficient? Can its effectiveness and efficiency be improved?

In general, the district should validate the assessments for each of their purposes, using strategies determined by the purpose. Validation documentation should:

Present the rationale for each recommended interpretation and use of test scores.

Summarize the evidence and theory bearing on the intended use or interpretation.

Validation evidence should include:

Expert review

Empirical studies

Relationships among different measures

Criterion group comparisons

Utility:
- Does the assessment provide useful information?
- Can results be used to improve learning and instruction?

Item analyses

Difficulty

Discrimination

Cognitive demands

Test properties

Difficulty

Inter-item correlations

Reliability

Evidence on differential validity

Investigation and elimination of irrelevant variance

In its June 30, 2006 deliverable, CRESST includes recommendations in each of these areas, based on ongoing discussions with the district and review of information on and data from the benchmark assessments.

Competencies Needed to Implement Recommendations

The initial recommendations in this document have an implied set of skills and knowledge that JPS district staff would need for effective implementation. It is possible that the staff members already possess the requisite knowledge base, which would limit the need for additional professional development or technical support. With that in mind, the implied skills include:

Content knowledge/expertise for each content area covered in the benchmark tests (particularly for test developers and senior reviewers; also necessary for teachers if teachers are being asked to make any curricular changes based on assessment findings).

Basic understanding of item design/qualities of good items (for test developers; see Appendix A for a summary of this information).

Knowledge of item scoring procedures including, if appropriate, use of rubrics (for test designers and any teachers expected to score student responses).

Understanding of classification systems of different levels of cognitive skills/demands (for test developers and senior reviewers; specific content dependant on the system selected by JPS for alignment analysis).

Basic understanding of properties of tests and their measurement. Specifically, this includes the skills needed to conduct statistical analyses of test quality such as inter-item correlations, reliability, and item analysis (for JPS testing/assessment office staff). *(Note: additional information about reliability analyses will be presented in the second deliverable to JPS.)*

Ability to conduct statistical studies of validity such as criterion group comparisons or evidence of differential validity (for JPS testing/assessment office staff). *(Note: additional information about reliability analyses will be presented in the second deliverable to JPS.)*

References

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

DeCoster, J. (1998). *Overview of Factor Analysis.* Retrieved February 14, 2006 from http://www.stat-help.com/notes.html

Hambleton, R. K. (1996). Advances in assessment methods. In Berliner, D. C., & Calfee, R. F. (Eds.), *Handbook of Educational Psychology* (pp. 245–276). New York: Macmillan.

Herman, J. L. & Baker, E. L. (2005). Making benchmark testing work for accountability and improvement: Quality matters. *Educational Leadership, 63*(3), 48-55.

Linn, R. L. (Ed.) (1993). *Educational measurement*. Phoenix, AZ: Oryx Press.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Porter, A. C. & Simthson, J. L. (2002, January). *Alignment of assessments, standards, and instruction using curriculum indicator data*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Rothman, R., Slattery, J.B., Vranek, J. L. & Resnick, L.B. (2002). *Benchmarking and alignment of standards and testing* (CRESST Technical Report 566). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education. Madison, WI: University of Wisconsin, Wisconsin Center for Educational Research.

Appendix A

Grade 2 Math Benchmark Item Analysis

		Benchmark Test						
Reporting Category	Benchmark /Item from MCT Blueprint	1st	2nd	3rd	4th	# items on benchmark tests	Total # of items per reporting category on benchmark test (% of total)	Total # items tested on MCT (% of total)
Patterns and Algebraic Thinking								
	1						11 (8.2%)	7 (15.5%)
	2		3			3		
	3	2	2			4		
	23	2	2			4		
Data Analysis								
	4	3	3	2	2	10	30 (22.4%)	7 (15.5%)
	5	1	1	1	1	4		
	6	3	3	3	3	12		
	7	1	1	1	1	4		
Measurement								
	2		1			1	30 (22.4%)	8 (17.8%)
	8			1	1	2		
	9			1	2	3		
	10			1	1	2		
	11			1	1	2		
	12			1	1	2		
	13			1	1	2		
	14			1	1	2		
	15			1	1	2		
	16			1	1	2		
	17			1	1	2		
	31		2	1		3		
	32		1			1		
	33				1	1		
	34				1	1		
	35				1	1		
	36		1			1		
Geometry								
	1				1	1	5 (3.7%)	6 (13.3%)
	18				1	1		
	19				1	1		
	20				1	1		
	21				1	1		
Number Sense								
	3	1	1			2	49 (36.6%)	17 (37.8%)
	22	2	2			4		
	23	1	1			2		
	24	1	1			2		
	25	1	1			2		
	26		2		1	3		
	27		2		1	3		
	28				1	1		
	29				1	1		
	30	1	1			2		
	37				1	1		
	38				1	1		
	39	1	1	1	1	4		
	40	1	1			2		
	41	1	2	2	1	6		
	42							
	43				1	1		
	44	2	1			3		
	45	1	1	1	1	4		
	46		1	1	2	4		

	47	1	2	1	1	5		
Total # of items		**32**	**40**	**24**	**38**			

Note. MCT = Mississippi Curriculum Test

Appendix B

Multiple Choice Item-writing Guidelines

David Niemi

Multiple-choice items are items in which a question (stem) and a set of possible answers (responses) are presented and the student is expected to choose the correct one. Incorrect responses are called distractors.

Two principles of assessment item design:

Students who know the content should be able to complete the item successfully. The language and format of the question should not mislead them.

Students who don't know the content should not be able to complete the item successfully (except for a slight possibility of guessing correctly).

Question wording

The most important and most obvious consideration in writing an assessment item is to make sure that the item measures the student's content knowledge and not some irrelevant skill. Science and math items, for example, should not require exceptional reading ability.

Information should be presented as clearly and concisely as possible so that students are not confused by what is being asked. At the same time, it should not be possible for students who do not know the content to get the item correct based on language cues.

Instructions should accompany each question or set of questions; for example:

Multiple Choice Item Instructions:
Circle the picture that shows six birds.
-OR-
Fill in the oval next to the correct answer.

Instructions in an item should be placed before graphics or other text.

Guidelines for multiple choice items

GUIDELINE 1

Avoid unusual or difficult vocabulary (unless that vocabulary is the focus of the item). Eliminate unnecessary wordiness, and avoid overly complex sentence structures.

Problematic Item:

Which of the following materials is known for having the property of combustibility, and in addition to possessing that property is one that will ignite in the least laborious fashion?

a. coal
b. gasoline
c. rubber
d. steel

GUIDELINE 2

In general, avoid fill-in-the blank and sentence completion formats. It is preferable to write questions.

Problematic Item:

Sacramento is the _____ city of California.

a. largest
b. westernmost
c. richest
d. capital

GUIDELINE 3

Make sure there is only one correct or best answer. The correct answer should be one that subject-area experts would agree is clearly the best.

Problematic Item:

Who is the best writer?

a. J. K. Rowling
b. C. S. Lewis
c. E. B. White
d. F. L. Baum

GUIDELINE 4

Avoid negative sentence structures, which tend to be confusing. Use "none of the above" and "all of the above" sparingly, especially in items for younger students.

Problematic Item:

Which statement about the set of whole numbers {1, 2, 3, . . . 20} is not true?

a. half are not even numbers.
b. more than 50% are not prime numbers.
c. 70% are not multiples of 3.
d. none of the above

GUIDELINE 5

Use plausible distractors. Watch for unintentional clues that make it possible to identify the correct response or to exclude incorrect response options.

Problematic Item:

What is the largest mammal?

a. huge blue whale
b. ant
c. redwood tree
d. small rock

GUIDELINE 6

Avoid absolute terms (e.g. always and never) and vague qualifiers (e.g. usually and generally).

Problematic Item:

Which of the following is true?

a. Cotton is never white.
b. Cotton is always white.
c. Cotton is usually white.
d. Cotton is generally used to make clothes.

GUIDELINE 7

In multiple choice questions, use 3 response options for kindergarten and first grade students, 4 for students in Grades 2-12.

GUIDELINE 8

Avoid repetitious language in response options.

Problematic Item:
How did Isaac Newton learn algebra?
a. Isaac Newton learned algebra by taking classes at Oxford University.
b. Isaac Newton learned algebra by studying algebra.
c. Isaac Newton learned algebra by taking classes at Cambridge University.
d. Isaac Newton learned algebra by inventing it.

Usually it is not necessary to use articles at the beginning of short responses; e.g., just use "boat" instead of "a boat" or "the boat" (where the question is, say, "Which of these is best for traveling across water?").

GUIDELINE 9

Avoid intentionally deceptive or "trick" questions.

Problematic Item:
Where did George Bush send his diplomats to negotiate with Napoleon?
a. Paris
b. London
c. Berlin
d. Nowhere

GUIDELINE 10

In general, there should only be one correct response to a multiple choice item.

GUIDELINE 11

Whenever possible, put measurement units in the stem rather than repeating them in the responses, e.g., "What is the length of this line in centimeters?"

GUIDELINE 12

Avoid using "you," e.g., "What should you use to measure volume?"

GUIDELINE 13

Randomize the order of response options, except for numbers, which can be arranged in ascending order.

GUIDELINE 14

These are guidelines, not laws. Item writers and reviewers will have to use judgment to determine (a) how to apply the guidelines in specific cases and (b) whether the two principles of assessment design have been observed or not.

Appendix C

Preliminary Analysis of the Jackson School District's First Term Algebra 1 Test, 2005-2006

Recommendations regarding reliability and validity analyses will be addressed in greater depth in the second deliverable to Jackson Public Schools (JPS). However, as an example of what some of this work might entail, we present an example of a preliminary test analysis. This example is based on findings from JPS's first term Algebra 1 test in 2005-06.

In the school year of 2005-06, 2,244 students at the Jackson School District took the 35-item Algebra 1 test. The students who took the test were enrolled in Grades 8-12, with the majority of students in Grades 8-10. There were 458 8th-graders, 780 students in 9th-grade, 850 in 10th-grade, 99 in 11th-grade, and 34 in 12th-grade. The gender breakdown for the student population was relatively even, with 54% female and 46% male. Of the total, 98% of the students were African Americans and about 90% received free or reduced fee lunch.

Descriptive Analysis

To describe how these students performed on the test, we started with descriptive analyses. First we calculated the percentages of students who scored correctly by item in order to give us an initial indication of item difficulty. Please see Charts A and B for the detailed results. Chart A has results on the first 18 items, and Chart B has the remaining 17 items. Note that 1.00 on the axis indicates that 100% of students answered an item correctly.

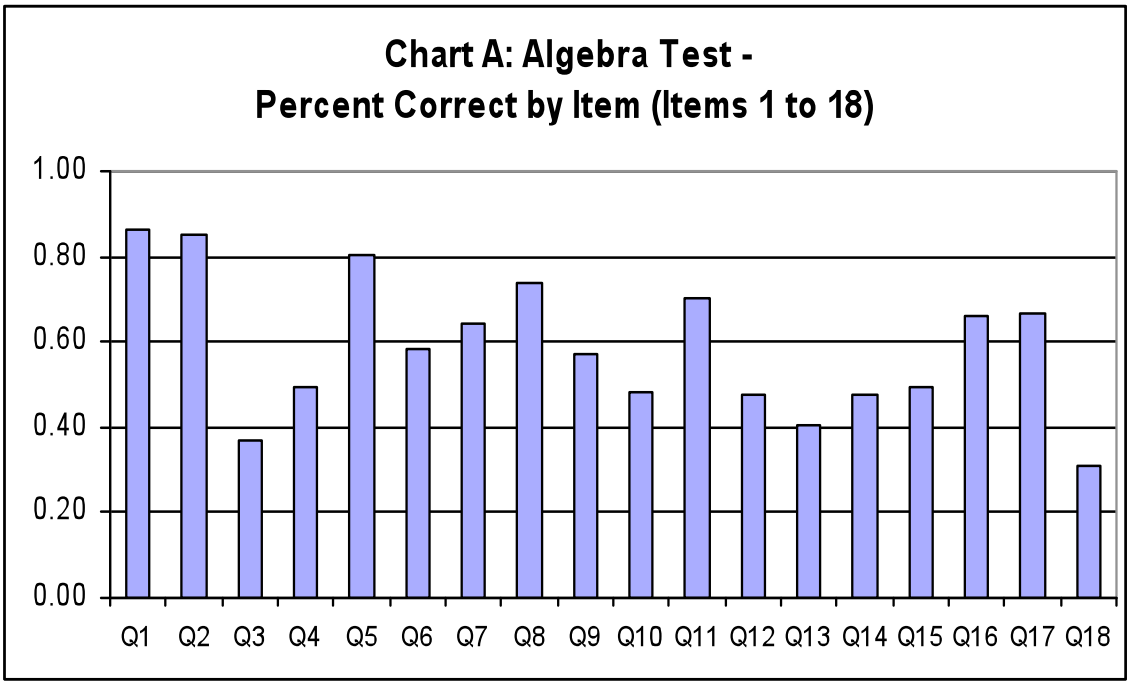
Chart A: Algebra Test -
Percent Correct by Item (Items 1 to 18)
1.00
0.80
0.60
0.40
0.20
0.00
Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16 Q17 Q18

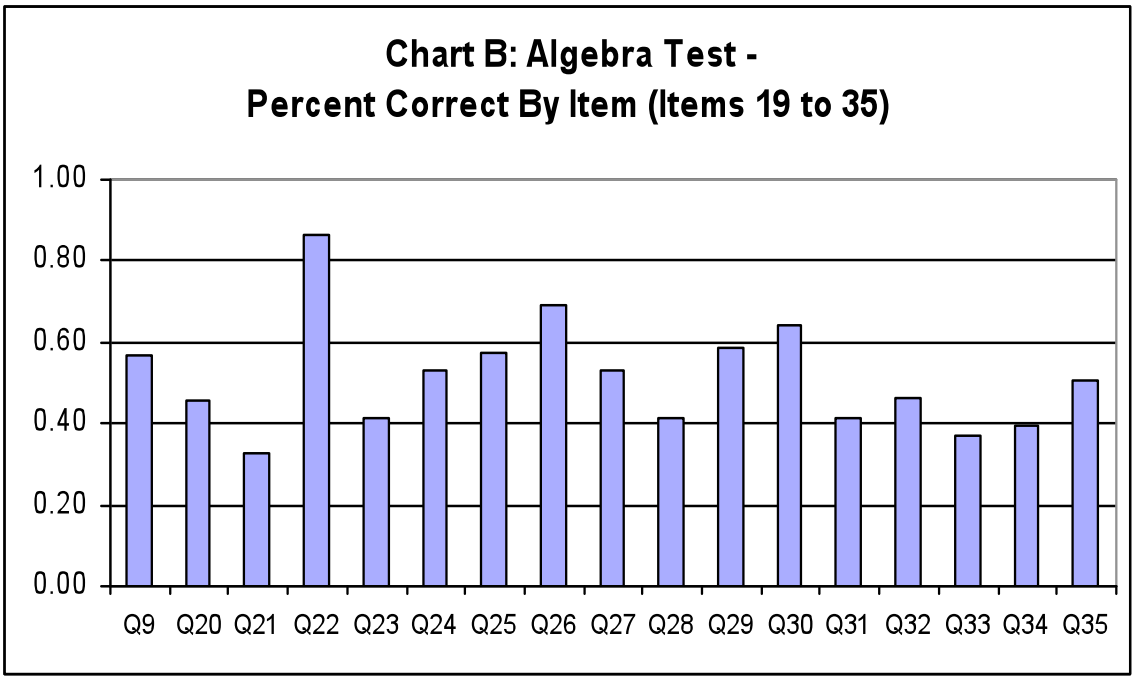
Chart B: Algebra Test -
Percent Correct By Item (Items 19 to 35)
1.00
0.80
0.60
0.40
0.20
0.00
Q9 Q20 Q21 Q22 Q23 Q24 Q25 Q26 Q27 Q28 Q29 Q30 Q31 Q32 Q33 Q34 Q35

As shown in Charts A and B, there is a wide variation in the percentages of students answering each item correctly across this 35-item test. Only about 31% of the students passed item 18, while 86% of the students passed item 22. Caution should be directed to items with passing rate less than 50%. These items should be examined in terms of content and phrasing to determine if they are addressing the specific knowledge/skills that they were designed to address. This process helps to ensure the *content validity* of the items. Items that, after analysis, are determined to have content validity problems should be deleted from the test or at least deleted from the calculation of the final scores for the students.

If the low pass-rate items are found to be valid in content, both students and teachers should be considered in investigating reasons for the pass rate. For example, it could be the content was not covered adequately in the classroom, or it could be the students did not master the content because they did not possess the requisite background knowledge or skills.

Besides analyzing the individual item passing rate for all the students who took the test, we also analyzed the individual item passing rate by various student background variables including gender, ethnicity, status in receiving free or reduced fee lunch, and grade level. There is no one specific pattern for the passing rates by grade level. In other words, some items seem to be of the same difficulty level for all grades, some items are more difficult for 8th-graders than for the students of higher grades, and sometimes items are more difficult for 12th-graders than for the students in lower grades.

The gender differences in passing items are relatively small; the maximum differences are 6% (for items 23, 25, 26, and 30). There were differences in students' passing rates by ethnicity, some as large as 23%. For example, for item 19, African American students had a passing rate of 47% while the other students had a passing rate of 70%. The differences in pass rate based on free/reduced lunch status are relatively minor when compared to the ethnic differences we found.

Correlational Analysis

The reliability coefficient for the 35-item test is .80, which is in the range generally considered to be acceptable for instrument reliability. In terms of investigating individual items, we found the Pearson correlation coefficients between each item and the mean score range from .19 to .49. After deleting 10 items with the lowest

correlations with the grand mean, the reliability coefficient for the remaining 25 items is .78, which is only slightly lower than the original one. The items dropped are items 31, 27, 5, 10, 3, 22, 1, 2, 8 and 11, whose correlations with the grand mean range from .19 to .31. It seems deleting these 10 items yields a more efficient test without reducing the general reliability of the test. These analyses suggested that the test internal reliability is reasonably high and that many items can be removed without seriously impairing that reliability.

Factor Analysis: Background

Factor analysis is a collection of methods used to determine the theoretical constructs that might be represented by a set of measures (items) in a given test, based on exploring the inter-correlation between these measures. The methodology of factor analysis is aimed to determine what common factors underlie a set of measurements and the nature of these factors.

There are two kinds of factor analysis: exploratory factor analysis and confirmatory factor analysis. With exploratory factor analysis, researchers examine analysis output to determine the number of factors underlying the set of measures. By allowing all items to load on all factors, exploratory factor analysis examines the factor structure and the internal reliability of the factors. In confirmatory factor analysis a specific, theory-based factor structure is defined prior to analysis indicating which items should load on which factor. Confirmatory factor analysis tests whether a specified set of factors is composed of specific items in a predicted way.

Researchers conduct factor analysis through examination of the pattern of correlations between the observed measures. Highly correlated (either positive or negative) measures usually represent the same factors, while relatively uncorrelated measures belong to different factors (DeCoster, 1998).

Factor Analysis: Results

We conducted a factor analysis with VARIMAX rotation on the 35 items. Factor analysis is applied to examine the pattern of correlations between the observed measures. Highly correlated (either positive or negative) measures usually represent the same factors, while relatively uncorrelated measures belong to different factors (DeCoster, 1998). Based on the district's assessment blueprints for the algebra test, these 35 items are supposed to represent 11 concepts labeled as 1A, 1B, 1C, 1D, 1E, 4A,

4B, 4C, 4F, 6C and 6F, with 3 or 4 items representing each concept, as is shown in the left two columns of the Factor Analysis Table below. The numbers in the table indicate the factor loadings of each item on each factor. Factor loadings are simply the correlations between the items and the underlying factors. With a higher factor loading value, we are more confident that the variables load on the factor.

The table shows that the current factor analysis identified 11 factors on the 35 items. However, most of the items actually load on the same factor, Factor 1. The items of Concepts 4A, 4C and 4F completely load on Factor 1. Factor 2 seems to embody 1E and part of 1C, factor 3 is related to concept 1D, Factor 4 has part of 1A and 6F, Factor 5 takes a big part of 6C, Factor 6 is related to 4C, Factor 7 is related to 4B, Factor 8 captures part of 1B and 6C, Factor 9 is related to 4B, and Factor 10 and Factor 11 are not significantly related to any of the concepts. From the factor analysis we see that the underlying concepts are highly correlated with each other and the items embody one general concept while retaining some of their own unique features. The findings suggest that the measures of the individual concepts on the test might not be as distinct as anticipated, and that the measure may be predominantly tapping into a more general algebra skill.

Factor Analysis Table

Blueprint Category	Question Number	Factors 1	2	3	4	5	6	7	8	9	10	11
1A	Q13	**0.40**	-0.24	0.34	-0.04	-0.03	-0.01	0.01	0.11	0.00	0.36	0.04
1A	Q14	**0.38**	0.06	-0.09	0.11	0.07	-0.12	0.02	-0.06	0.37	0.02	-0.14
1A	Q20	0.37	0.16	0.06	**0.38**	-0.03	-0.10	0.27	-0.03	0.17	0.12	-0.29
1A	Q27	0.19	0.17	-0.14	***0.57***	-0.21	0.14	0.15	-0.04	0.31	0.07	0.13
1B	Q6	**0.39**	-0.02	-0.10	0.20	0.18	0.04	0.01	0.16	-0.19	-0.26	0.04
1B	Q10	0.21	0.18	-0.17	0.27	0.13	-0.10	-0.06	**0.47**	-0.34	-0.02	0.34
1B	Q30	***0.52***	-0.09	-0.09	0.00	0.09	0.07	0.17	-0.03	-0.07	-0.29	-0.07
1C	Q8	0.30	**0.42**	0.11	-0.13	0.05	-0.15	0.01	-0.22	-0.06	-0.22	0.18
1C	Q15	**0.40**	0.14	0.19	0.01	-0.22	-0.07	0.10	0.12	0.05	-0.23	0.06
1C	Q21	**0.34**	-0.10	0.25	-0.02	-0.37	-0.12	-0.35	0.13	0.11	0.01	0.14
1C	Q22	0.27	***0.50***	0.20	-0.20	0.15	0.10	-0.03	0.12	-0.09	0.06	-0.18
1D	Q16	**0.42**	0.25	0.12	0.03	-0.02	-0.25	0.16	-0.29	-0.11	-0.13	-0.09
1D	Q18	0.31	-0.23	0.31	0.16	-0.06	**0.34**	-0.14	-0.11	-0.28	-0.14	0.11
1D	Q23	**0.35**	-0.20	0.34	0.11	-0.12	0.30	-0.24	-0.14	0.04	-0.30	-0.11
1E	Q1	0.28	**0.45**	0.20	-0.25	0.04	0.19	0.06	0.13	0.05	0.16	0.12
1E	Q5	0.22	**0.39**	0.20	-0.17	0.12	-0.02	-0.22	-0.01	0.18	0.06	0.05
4A	Q2	0.29	-0.08	0.01	0.03	**0.45**	0.32	-0.01	-0.15	0.16	0.08	0.18
4A	Q3	0.26	-0.12	0.10	0.13	**0.36**	-0.07	0.09	-0.43	-0.08	0.31	0.28
4A	Q7	**0.43**	-0.08	0.07	-0.08	0.25	0.07	0.02	-0.17	0.00	-0.15	-0.15
4B	Q9	**0.44**	-0.20	-0.10	-0.04	0.14	0.01	0.08	0.05	-0.14	0.23	-0.07
4B	Q11	0.29	0.20	-0.10	**0.42**	0.09	0.12	-0.27	0.07	0.04	0.05	0.09
4C	Q4	**0.42**	-0.18	0.24	-0.06	-0.01	-0.13	-0.15	0.10	-0.05	0.22	-0.02
4C	Q12	**0.49**	-0.11	-0.14	0.00	0.14	-0.29	-0.10	0.17	-0.05	0.00	0.03
4C	Q19	**0.39**	-0.15	0.28	0.10	-0.15	-0.21	0.21	0.17	0.00	0.21	-0.10
4F	Q31	0.14	-0.02	0.13	-0.17	-0.16	0.33	***0.64***	0.16	-0.09	-0.01	0.17
4F	Q32	**0.48**	-0.21	-0.26	-0.24	0.11	0.01	0.07	0.12	0.12	-0.05	0.16
4F	Q34	0.37	-0.09	-0.09	-0.21	0.00	-0.06	-0.09	0.08	**0.42**	-0.10	0.32
6C	Q17	**0.36**	0.07	-0.31	-0.14	0.00	0.04	-0.11	0.00	-0.22	0.14	-0.26
6C	Q24	**0.43**	0.10	-0.04	0.03	-0.32	-0.01	-0.08	-0.20	-0.29	0.07	0.01
6C	Q26	**0.40**	0.02	-0.25	-0.09	-0.07	0.30	-0.17	-0.02	0.04	0.21	-0.27
6C	Q29	0.30	0.08	-0.30	-0.06	-0.28	**0.40**	0.00	0.04	0.07	0.09	0.01
6F	Q28	**0.39**	0.02	-0.27	-0.01	-0.25	-0.18	0.04	-0.31	-0.20	0.05	0.14
6F	Q33	**0.42**	-0.23	-0.03	-0.16	-0.05	-0.16	0.10	-0.03	0.18	-0.18	0.00
6F	Q35	**0.44**	0.00	-0.34	-0.18	-0.18	-0.04	-0.03	-0.09	0.01	-0.01	0.05