

THREE FACETS OF FORMATIVE ASSESSMENT

How to Revolutionize (and actually use)
Locally Developed Tests

Dan Mason, M.S.
Mark Moulton, Ph.D.
Dale Russell, Ed.D.
Diana Wilmot, Ph.D.

Spring 2009

Introduction

This guide is intended for educators who are facing increasing requirements to show their instructional programs' work. Teachers and administrators occasionally place students in classes according to performance data that are not accurate or timely. Then, teachers are expected to show results, often using inadequate tools. There is a better way!

Student placement can be based on valid and reliable data directly related to the written, taught, and tested curriculum. Teaching expressly tied to the intended curriculum can yield results that are communicated consistently. Perhaps most important, teachers get engaged when they examine their own teaching effectiveness using clearly presented and understandable assessment results. Student proficiency can be more than comparisons of grades based on the percent correct score of the classroom tests. These benefits, plus an added value of data that reliably predict high stakes test outcomes, are topics of this guide.

The approach we refer to as the “Three Facets of Formative Assessment” rests with using well-designed locally developed tests to not only enable the teacher to plan and implement sound instruction, but also to guide policy discussions about delivering useful, high quality assessment results that students, parents, and teachers deserve.

The following graphic summarizes the concepts discussed in this document. It was developed in collaboration among the writers to share what we believe is a helpful conceptual relationship involved in examining data from locally developed tests. It begins with a rather widespread practice of reviewing percent of students that correctly answer an item and possible reasons for selecting wrong answers or distractors. The scheme then moves to examining the data as items are found to interact in difficulty with more-to-less proficient students. Finally, the scheme leads to exploring use of formative tests to predict student proficiency on such high stakes tests as state proficiency exams.

Section One of this guide begins the process of developing good formative assessments. Section Two unpacks the tasks included in analyzing test items and assisting teachers analyze reasons students may have performed as they did on each item. Section Two also introduces an application to using test items with students to engage students, as well as teachers, in reflection on the thought process utilized by students to solve problems. Section Three discusses how to use formative assessments as “mini high stakes tests” and accurately predict performance on the high stakes test through applying psychometric procedures. Section Four introduces some policy implications that should be considered when evaluating present assessment systems as practiced in California.

Our special thanks and appreciation go to Peggy Stull for her valuable assistance.

Three Facets of Formative Assessment				
<i>Facets</i>		I	II	III
<i>Activity</i>		Attractive Distractor Analysis	Cluster analysis	CST metric scores
<i>Interpretation</i>		Misconceptions	Mental operations, cognitive processes, how students learn	Predicted performance on CST
<i>Implications</i>		Pacing guide implementation	Instructional strategy analysis/evaluation	Consensus finding
<i>Application/use</i>		Instructional Refinement	Organization of standards	Monitor progress towards CST
<i>Available DATA</i>		Raw Score	Scaled Scores	
4 Levels of Data Analysis: Driving Questions	<i>Item Level</i>	What do attractive distractors in the most difficult items tell us about student misconceptions?	What do the most difficult items for students have in common?	
	<i>Standards Level</i>	How are the most difficult items reflected across standards? What are the most difficult standards for students?	How can I organize the State standards in a way that reflects a developmental progression of student learning?	
	<i>Classroom Level</i>	How can I develop lesson plans to address: student misconceptions? most difficult standards for students?	How can I develop lesson plans that reflect the developmental progression in which students learn the content?	How can I identify students in my class that are struggling to meet proficiency on the CST?
	<i>Grade Level</i>	How can I develop grade level instructional goals related to student performance?	How can I develop curriculum to help students attain better understanding of the content?	How does student performance on the Benchmark predict performance on the CST? How can we identify our "bubble kids"?

TABLE OF CONTENTS

Introduction	i
Three Facets of Formative Assessment	ii
SECTION I: WRITING TEST ITEMS FOR FORMATIVE ASSESSMENTS	3
Formative Assessments Development Chart	
Writing a Good Question	
Things to Consider – Generally	
Design tips	
Bias guidelines	
Guidelines for Specific Item Types	
Completion/Short Answer	
Performance	
Essay	
Matching	
Multiple choice	
True-False	
Designing a “Blueprint for an Assessment that is incorporated into Progress Monitoring using a Pacing Guide”	
Think Aloud Protocol Script	
Acceptable and Unacceptable Questions	
Sample Questions Released from Other States	
SECTION II: THREE FACETS OF ANALYZING FORMATIVE ASSESSMENTS	35
Facet I: Inform Classroom Instruction: Identify student misconceptions	
Phase 1: Activating and Engaging — Making predictions and assumptions	
Phase 2: Exploring and discovering—Analyzing the data for “attractive distractors”	
Phase 3: Organizing and Integration—Establishing next steps to undo misconceptions	
Facet II: Inform Curricular Mapping: Recognize students’ development as a trajectory	
Compare students proficiency with item difficulty	
Identify content in students’ target “zone”	
Validate and refine development of model of student learning	
The Algebra I Progress Maps	
Validity Evidence based on internal structure	
Item fit analysis	
Convergent Evidence	
Validity Evidence based on response processes	
Discussion	
Facet III: Inform Programmatic Intervention: Understand students’ needs	
Moving beyond a conjecture of which students need remediation	

SECTION III: USING FORMATIVE ASSESSMENTS TO PREDICT PERFORMANCE	65
How Benchmark Exams Can Be Turned into Mini-CSTs	
Why Local Benchmark Exams	
Difficulties with Local Benchmark Exams	
The Benchmark Scaling Method Used by Educational Data Systems	
Applying Benchmark Scaling Methodology	
Validation	
Conclusion	
SECTION IV: BRINGING PRACTICE TO POLICY	79
PSAA and API	
No Child Left Behind and AYP	
Limits of CST scores	
FIGURES	85
TABLES	86
REFERENCES	87

SECTION I: WRITING TEST ITEMS FOR FORMATIVE ASSESSMENTS

Formative assessments serve the roles of guiding instruction and monitoring student proficiency gains. They are not intended to be high stakes but serve instructional purposes best when prediction to summative assessment follows from their use. It is helpful to think about formative assessment as being either FOR learning and AS learning. While summative assessments provide useful evaluative and policy level information, teachers and students must know whether the taught curriculum is learned. When the taught curriculum is not learned as well as desired “FORmative” assessments must guide the teacher and student toward interventions that succeed in closing that gap. Students often benefit by studying actual items to deconstruct the misunderstandings contained in the item as well as the correct response so they can see for themselves why they missed the item. This approach we consider assessment AS learning. In each instance, whether it is assessment OF, FOR, or AS learning, test items are developed following the same steps. The rigor required developing test items that are used in summative tests (OF learning) and progress monitoring tests (FOR learning) must be maintained at a high level. Items allowed for practice or student independent analysis, AS learning need not be so rigorous but must present logical analysis that supports use in cognitive labs or think aloud strategies to improve student analytical, and test taking skill, development.

In this first section we will focus on writing test items. Since it is our purpose to place the developed items into an item bank to deliver to teachers through a mechanized system we will discuss item writing in context of an item bank deployment.

With this in mind, our purposes are to create an item bank for two central uses: 1) Items that can be used for purposes of formatting local assessments used at preset increments for student progress monitoring against curriculum pacing guides commonly found in districts; 2) Items can be drawn from the bank and formatted into local assessments “on demand” to meet the requests for uses by individual teachers or as desired by schools and/or districts.

A prerequisite to meeting these objectives is to format the tests into forms that closely resemble the California Standards Tests or the California High School Exit Examination.

A secondary purpose of the guide is to describe steps to devise performance based tests with scoring rubrics and efficient reporting schemes.

Developing items for an item bank can be done by individuals or writing teams. We will proceed as if writers are organized in grade level teams.

Grade level teams will create (or review) grade level assessments while observing the following concepts and principles.

1. Validity and reliability are the key concerns with item development. State law, and sound instructional practice, requires tests to be valid and reliable.

2. Review Pacing Chart, Scope and Sequence of instruction, textbooks, end of unit tests and related curriculum content to see what is being taught and therefore what should be tested. Agree on what STANDARDS will be assessed.
3. Identify essential standards to test at each grading period such as quarter, semester, or trimester.
4. Develop item specifications for each standard to be assessed.
5. Assure that all items are mapped to standards.
6. Develop or select questions that measure selected standards.
7. Evaluate item quality. Learn what makes an item acceptable or unacceptable.
8. Select or write items to assure that each assessed standard has five to six items per standard.
9. Have a sufficient number of items to assess each standard. Interpretation of student proficiency on each assessed standard will be based on viewing the data collectively for item sets. If there are too few items it is difficult to have confidence in results.

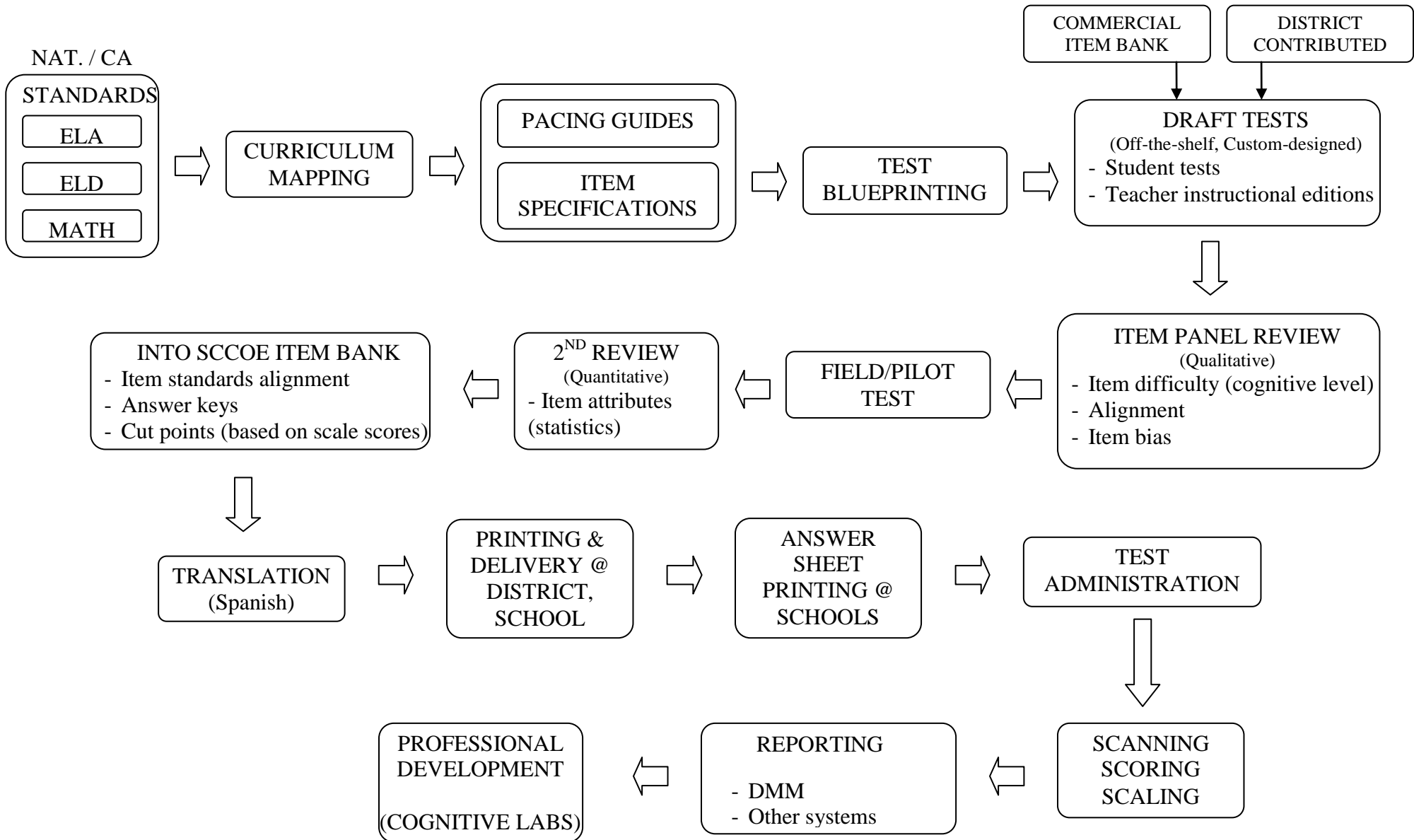
It is helpful for test developers to understand the ways test results will be accessed and presented for use by various consumers. Some questions to consider when examining the data system for accessing results are outlined here:

Determine features of the data management system which will be used by educators to see test results.

1. How will student responses be recorded? How will recorded responses be summarized and distributed to teachers? Will teachers hand score tests or use a scanning solution?
 - a. Review the mechanics of scanning, scoring, and reporting.
 - b. What equipment needs must be met before scanning will work?
 - c. Select staff to format the tests, provide oversight for printing, validate the answer keys, and agree on report formats required.
2. Will teachers meet in teams or individually analyze data and interpret results?

The process followed to develop formative assessments is outlined in the flow chart on the next page.

FORMATIVE ASSESSMENTS DEVELOPMENT CHART



Writing a Good Question

Item writers must be mindful of technical conventions to observe (AERA/APA/NCME, 1999). Some of the more important ones have been outlined for ready reference in this section. Item specifications are used to assure uniformity of item development. It is necessary to have several, usually at least five, items to measure student proficiency on each standard. By first developing item specifications writers are more able to write items that are similar and have greater likelihood of fitting accurately into the constructed test. Since most of the items that will be developed for the item bank will be multiple choice, examples of item specifications will be shown in that part of this section of the guide.

Things to Consider — Generally

- It is imperative that the item writers be content experts. “The item must focus the attention of the examinee on the principle or construct upon which the item is based.” (Academic Technology Services, Michigan State University. n/d). Well-written items will permit an analysis of test results that reveal strength of student knowledge as well as the misperceptions that lead to an incorrect answer choice. High level content knowledge of item writers is the best assurance that these criteria are met.
- Before you begin, meet as a group to determine formatting rules. This will ensure consistency of style, format, text, and graphics within items and subject areas.
- Group like item types together. Respondents should be able to answer items of one type without shifting back and forth from one type of task to another. Grouping also makes it easier for administration directions to be clear and specific.
- Make sure the question you’re writing matches the standard or skill description.
- Keep the correct answer and the distractors about the same length, or if the distractors and answers are in two different forms, use pairs of similar items.
- Be aware of obvious giveaways like having the correct answer be a positive statement and the distractors negative statements, having the correct answer be clearly longer or shorter than the distractors, or having the correct answer be a whole number when the distractors are fractions.
- Include as much information in the question as possible so the answers don’t have to repeat information.
- Avoid long sentences as answer choices.
- Use full word names when possible (miles, inches, etc.).
- Questions should be at an appropriate reading level for the grade level for which they are written.
- Questions topics should be relevant to the grade level and contain appropriate content. Avoid writing questions/passages about people who are still living.
- Make sure all answers are plausible mistakes for the given grade level and skill. The goal isn’t to trick learners; it’s to present foils that are plausible to learners who haven’t learned the material.
- Items shouldn’t contain any offensive material.
- Go to: <http://www.babycenter.com/babynames/> for a good list of names to use in questions.
- Use lots of space between instructions and questions. Use plenty of space around graphics. A good rule is to use a double return between all instructions and questions and before and after a graphic in a question.

- It's a good idea to bring attention to words that could cause the reader to misunderstand a question such as *not*, *best*, *most likely*, *least*, etc.
- Avoid using trademarked names such as Kleenex, Adidas, Jell-O (or jello), Toyota, etc. Use: <http://www.ascendercorp.com/about/trademarks/> to see if a word is trademarked. This site has lists of trademarked words arranged alphabetically. Wikipedia also is an excellent source for searching trademark information.

Design Tips

- Relevant graphics are extremely helpful; use them whenever possible.
- Unclear graphics can hinder learning through distraction, disruption, and/or seduction.
- Put corresponding words and graphics together.
- Be consistent with style, format, text, graphics, etc.
- Avoid adding extraneous words.
- Use vocabulary that is consistent with the intended grade level of the item.

Bias Guidelines

- Avoid gender stereotyping (females cooking, females cleaning).
- Avoid ethnic bias such as referring to various races or nationalities engaging in stereotypical activities.
- Avoid continuing any stereotype.
- Use common ethnic names in lower grade levels instead of more difficult ones so names don't provide unnecessary distraction or add to the difficulty of an item.

Guidelines for Specific Item Types

Some guidelines to consider when developing a test are related to time to complete each item of different types as well as advantages and disadvantages of each item type. An exhaustive set of guidelines for each item type is beyond the scope of this document. Guidelines for selected item types are briefly presented below.

Completion/Short Answer

A completion or short answer test is one that requires the student to create a response in the form of one or more words or phrases. These items require students to supply a response rather than select an answer from provided options. They are frequently used for recall of information or problem solving in math or science when a correct solution or calculation is possible. A short answer question is designed with only one correct or clearly "best" answer. A common type of short answer question is one where the question is in the form of an incomplete sentence. The student must "complete" the sentence by filling in the missing word or phrase. They do permit a broad sampling of material but usually require hand scoring and are limited to lower cognitive levels.

An example of a completion item is:

There are _____ inches in a foot.

Examples of short answer items are:

How many inches are there in a foot? _____

Define "Vegetarian". _____

Some guidelines for writing completion and short answer items are:

1. The requested answer should be brief and specific.
2. Answers should be in a consistent location to avoid scoring errors (e.g. within the body of the item or on the right hand margin).
3. There should be only one blank in the item unless the answer requires terms that are part of a series.
4. The wording and grammar should not provide clues to the answer (“a/an” “is/are”)
5. If the answer is a number, indicate the unit of measurement (pounds, cents, dollars, etc.) and the degree of specificity (three decimal places) required.
6. Avoid response queues such as long and short blanks.

Performance

Performance assessment is a form of testing that requires students to perform a task rather than select an answer from a ready-made list or provide a short, limited response to a question. This type of assessment is also known as alternative or authentic assessment.

Examples of performance assessment items are: Ask a student to explain historical events, generate scientific hypotheses, develop proofs of math problems, converse in a foreign language, or illustrate a scientific principle involved in a given context.

To score performance items raters judge the quality of the student’s work based on an agreed-upon set of criteria often referred to as a rubric. The rubric provides a single score value that summarizes the agreed performance level of the student work product. When developing the rubric it is essential to describe what the task entails and the standards that will be used to evaluate performance.

Following are some methods that have been used successfully to assess performance:

- Open-ended or extended response exercises are questions or other prompts that require students to explore a topic. Students might be asked to describe their observations from a science experiment, present arguments defending an action taken in history, advocate for or against a position or proposition or similar task. For example: What would Abraham Lincoln argue were the causes of the Civil War?
- Extended tasks are assignments that require sustained attention in a single work area and are carried out over several hours or longer. Such tasks could include drafting, reviewing, and revising a poem; conducting and explaining the results of a science experiment on photosynthesis; or even painting a car in auto shop.
- Portfolios are selected collections of a variety of performance-based work. A portfolio might include a student’s “best pieces” and the student’s evaluation of their strengths and weaknesses. The portfolio may also contain some “works in progress” that illustrate improvements made over time.

Proponents of performance assessments contend that they require students to actively demonstrate what they know and are therefore a more valid indicator of students’ knowledge and abilities. They point to such things as the difference between answering multiple choice questions on how to make an oral presentation and actually making an oral presentation.

Proponents also contend that performance assessments results provide impetus for improving instruction while increasing students critical self-reflection. When preparing students to work on performance tasks teachers need a careful description of the elements of good performance that allows students to judge their own work as they proceed. Performance tasks must be inherently instructional and actively engage students in worthwhile learning activities.

Performance assessment requires a greater expense of time, planning and thought from students and teachers. Teachers must spend more time planning and more time coaching for this type of assessment to have optimal value. Users also need to pay close attention to technical and equity issues to ensure that the assessments are fair to all students.

Essay

Essay items are a kind of performance assessment since the respondent must complete a task to receive credit. Prompts are provided and the respondent must write a narrative that conforms to the requirements of each specific prompt. Prompts may require the respondent to develop a narrative using one of several styles, referred to as genres. Examples of genres are: autobiographical narrative, summary, information report, and response to literature. Good essay questions are demanding to develop, administer, and score. For example, an essay item must include a prompt that clearly identifies the genre assessed, have unambiguous directions for administration and responding, be accompanied by a clear rubric with authentic examples that illustrate each score value identified in the rubric, and a means of presenting the score in context of the proficiency standards attained.

Consider the Grade 4 California Writing Standards Test that was administered in 2006 and subsequently released. The item included directions, scoring criteria, prompt, space for planning the narrative, and space for the actual narrative the student will submit for scoring.

ACTUAL BOOKLET IS NOT SHOWN

Writing Prompt and Response Booklet

Narrative Writing Task

Directions:

- In this writing test, you will respond to the writing task on the following pages.
- You will have time to plan your response and write a first draft with edits.
- Only what you write on the lined pages in this booklet will be scored.
- Use only a No. 2 pencil to write your response.

Scoring:

Your writing will be scored on how well you

- include a beginning, a middle and an end;
- use details; and
- use correct grammar, spelling, punctuation, and capitalization.

BOOKLET CONTINUES...

Read the following writing task. You must write a narrative about this topic.

Writing a Narrative

Imagine that you are asked to keep an elephant for a week. Write a story about your unusual experiences with your elephant.

When you write about this experience, remember

- to include a beginning, a middle, and an end;
- to use details to describe the experience; and
- to use correct grammar, spelling, punctuation, and capitalization.

SPACE IS THEN PROVIDED FOR THE PLANNING USING BLANK PAPER, AND THE ACTUAL NARRATIVE USING LINED PAPER. THE DIRECTIONS ARE REPEATED IMMEDIATELY PRIOR TO THE LINED PAPER PORTION OF THE BOOKLET.

Essay questions are especially suited for assessing at:

- Application, synthesis, and evaluation levels

Types of essay questions:

- Extended response – synthesis and evaluation levels that have open ended form
- Restricted response – more consistent scoring, outlines parameters of responses

Advantages of essay questions:

- Students are less likely to guess
- Relatively easy to construct
- Requires more in depth knowledge of most subjects
- Allows students to demonstrate ability to organize knowledge, express opinions and show originality

Disadvantages of essay questions:

- May be flawed by subjective scoring
- Scoring requires calibration of scorers and monitoring for consistency
- Time consuming to score

Tips for writing good essay items:

- Provide ample time for planning and writing
- Sample from among available genres
- Use clear definitive directions that include the specific verb for the required cognitive level being assessed: compare, analyze, evaluate, etc.
- Use a consistent scoring rubric with model “anchor” papers that have been scored by curriculum experts with high reliability in the scoring process.
- Score one question at a time and all at the same time.

Matching

Matching test items, along with true-false and multiple choice items are selection items. They are specialized for use when measuring the student's ability to identify the relationship between a set of similar items, each of which has two components, such as words and their definitions, symbols and their meanings, dates and events, people and their accomplishments, etc. Of the two objectives listed below, only the second one is appropriate for a matching item

Objective A: Students will be able to explain the process of photosynthesis.

Objective B: Students will be able to identify primary characters in novels they read.

In measuring accomplishment of Objective A, the question would probably be one calling for the student to write a response. In contrast, Objective B states that the students will be able to "identify" primary characters. This implies some type of selection question in which the answers are provided, and the task of the student is recognition. The rest of the objective (primary characters in novels they read) indicates a series of novels, each with its respective primary character.

One matching item can replace several true-false or short answer items (and require less reading for the students). Matching items are generally easy to write and score when the test content and objectives are suitable for matching questions. Possible difficulties in using matching items may arise due to poor student handwriting or printing, or students' being able to guess correct answers through the process of elimination.

In developing matching items, there are two columns of material (Example 1). The items in the column on the left (Column A) are usually called premises and assigned numbers (1, 2, 3, etc.). Those in the column on the right (Column B) are called responses and designated by capital letters, as in Example 1. Capital letters are used rather than lower case letters in case some students have reading problems. Also there are apt to be fewer problems in scoring the student's handwritten responses if capital letters are used.

- 1. Directions: On the line next to each children's book in Column A print the letter of the animal or insect in Column B that is a main character in that book. Each animal or insect in Column B can be used only once.**

Example 1

Column A	Column B
_____ 1. Charlotte's Web	A. Bear
_____ 2. Winnie the Pooh	B. Chimpanzee
_____ 3. Black Beauty	C. Cricket
_____ 4. Tarzan	D. Deer
_____ 5. Pinocchio	E. Horse
_____ 6. Bambi	F. Pig

The student reads a premise (Column A) and finds the correct response from among those in Column B. The student then prints the letter of the correct response in the blank beside the premise in Column A. An alternative is to have the student draw a line from the correct response to the premise, but this is more time consuming to score.

In Example 1, the student only has to know five of the six answers to get them all correct. Since each animal in Column B can be used only once, the one remaining after the five known answers have been recorded is the answer for the sixth premise. One way to reduce the possibility of guessing correct answers is to list a larger number of responses (Column B) than premises (Column A), as is done in Example 2.

Example 2

Column A	Column B
1. Charlotte's Web	A. Bear
2. Winnie the Pooh	B. Chimpanzee
3. Black Beauty	C. Cricket
4. Tarzan	D. Deer
5. Pinocchio	E. Horse
6. Bambi	F. Mouse
	G. Pig

Some writers suggest there be no more than five to eight premises (Column A) in one set. For each premise, the student has to read through the entire list of responses (or those still unused) to find the matching response. For this reason, the shorter elements should be in Column B, rather than Column A to minimize the amount of reading needed for each item. Although there is little difference in the length of items in the two columns in Examples 1 and 2, note the improvement in Example 2b when the items in the two columns in Example 2a are reversed.

2a. Directions: On the line next to each description in Column A, place the letter of the president in Column B whom it describes. Answers in Column B may be used only once.

Column A	Column B
____ 1. Jimmy Carter	A. Our first President
____ 2. Abraham Lincoln	B. Resigned from the office of president
____ 3. Richard Nixon	C. Was well known for his association with humanitarian causes after leaving office
____ 4. George Washington	D. Was a movie star and a state governor before being elected president
____ 5. Ronald Reagan	E. Was assassinated while in office

2b. Directions: On the line next to each description in Column A, place the letter of the president in Column B whom it describes. Answers in Column B may be used only once.

Column A	Column B
_____ A. Our first president	1. Jimmy Carter
_____ B. Resigned from the office of president	2. Abraham Lincoln
_____ C. Was well known for his association with humanitarian causes after leaving office	3. Richard Nixon
_____ D. Was a movie star and a state governor before being elected president	4. Ronald Reagan
_____ E. Was assassinated while in office	5. George Washington

Responses (Column B) should be listed in logical order if there is one (chronological, by size, etc.). If there is no apparent order, the responses should be listed alphabetically. Premises (Column A) should NOT be listed in the same order as the responses, however, as in Example 3.

3. Directions: On the line next to each author in Column A, place the letter of the type of writing in Column B for which the author is best known. Answers in Column B may be used only once.

Column A	Column B
_____ 1. James Michener	A. History
_____ 2. Stephen King	B. Horror
_____ 3. Erma Bombeck	C. Humor
_____ 4. Agatha Christie	D. Mystery
_____ 5. Walt Whitman	E. Poetry
_____ 6. Danielle Steele	F. Romance
_____ 7. Isaac Asimov	G. Science Fiction

As previously mentioned, there should be a larger number of responses (Column B) than premises (Column A) to reduce the possibility of guessing correct answers. Another way to decrease the possibility of guessing is to allow responses to be used more than once. Directions to the students should be very clear about the use of responses. Example 4 utilizes both of these techniques: more responses than premises, and allowing responses to be used more than once.

4. Directions: On the line next to each author in Column A, place the letter of the type of writing in Column B for which the author is best known. Answers in Column B may be used once, more than once, or not at all.

Column A	Column B
_____ 1. Agatha Christie	A. History
_____ 2. Isaac Asimov	B. Horror
_____ 3. Erma Bombeck	C. Humor
_____ 4. Walt Whitman	D. Mystery
_____ 5. Stephen King	E. Poetry
_____ 6. James Michener	F. Science Fiction
	G. Tragedy

A difficulty sometimes arises in finding sufficient homogeneous material. In Example 5, the content can readily be sorted into two categories with 3 items each. In other words, only three questions relate to inventors, so the student only has to know two of them to get all three correct. The same is true for the three artists.

5. Directions: On the line next to each accomplishment in Column A print the letter of the person in Column B who is associated with that accomplishment. Each name in Column B can be used only once.

Column A	Column B
_____ 1. Discovered electricity	A. Thomas Edison
_____ 2. Famous for composing waltz music	B. Benjamin Franklin
_____ 3. Composed marches, such as the Stars & Stripes Forever	C. George Gershwin
_____ 4. Invented the telephone	D. Louis Pasteur
_____ 5. Wrote musical scores for Broadway shows	E. John Phillip Sousa
	F. Johann Strauss

Variation:

Example 6 shows one variation using a short list of answers, each with a capital letter designation, positioned above a set of items. Each question can be answered by using one (or sometimes more than one if this is specified in the directions) of the answers in the “key” which you have provided. The letter designating the correct response is printed in the blank beside the item.

6. Directions: Listed below are some objectives. In the blank beside each objective, specify the most appropriate type of assessment by placing the letter of the assessment type in the blank beside the objective.

Objectives	Assessment Types
_____ 1. Students will be able to construct a fluxty.	A. Essay
_____ 2. Students will know the six rules for effluding ixons.	B. Performance Assessment (portfolio, presentation, project, etc.)
_____ 3. Students will be able to explain to parents how their fluxty operates.	C. Traditional paper and pencil test (True/False, multiple choice, etc.)
_____ 4. Students will be able to evaluate the advantages and disadvantages of the various types of zibixs.	

General guidance:

- Check your objectives to make sure this type of question is appropriate.
- Include more responses than premises **OR** allow responses to be used more than once.

- Put the items with more words in Column A.
- Arrange items in Column B in either a logical or natural order or alphabetically if there is no apparent organizational basis.
- Use numbers to identify items in Column A, capital letters to identify responses in Column B.
- Correct answers should not be obvious to those who don't know the content being taught.
- Do **NOT** list premises in the same order as responses, and there should **NOT** be a pattern in the correct answers.
- There should **NOT** be keywords appearing in both a premise and response providing a clue to the correct answer.
- The items should all be part of a common set. It should **NOT** be possible to subdivide the premises and responses into two or more discrete subsets.
- All of the responses and premises for a matching item should appear on the same page.
- Directions to the students should explain how many times responses can be used.

Multiple-choice

Multiple-choice items consist of a stem that defines the question and answer options from which the correct answer is selected. It is helpful for item writers to review a checklist of item qualities to assist in keeping on track as items are written.

The Stem:

- If the stem is a question, start it with an interrogative word.
- Do not force the stem into the form of a direct question if an incomplete statement is more appropriate.
- Clearly define the question.
- Include as much of the item as possible in the stem leaving less for answer options.
- Avoid leaving blanks for completion in the beginning or middle.
- Use clear and simple language.
- If the item is measuring vocabulary, the highest level of language used in the stem should be below that considered appropriate for the grade or performance level being tested.
- Avoid negatives or double negatives; if a negative is used clearly emphasize it (e.g. capitalize all letters of the negative word).

The Answer:

- There should be only one correct answer to an item.
- Options should be grammatically consistent with the stem.
- Options should be parallel in form.
- Distractors or foils should be plausible and attractive to the examinee who does not know the correct answer.
- Write at least three distractors for every question.
- Do not force a fourth or fifth choice into an item which logically can have only three choices.
- Make all options independent of each other.
- Choices should be in logical order unless the order reveals the answer.
- Numerical responses should be from smallest to largest number, or the reverse.

- Single-word answers should be alphabetized unless there is logic for another order, such as months of the year.
- Lengthy responses should be arranged in order of their length.
- Choices that are identical with names of things on a graph should be ordered as they are on the graph.
- Options should be independent and mutually exclusive.
- Symbols used to identify alternatives should be used in a way that they cannot be confused with the content of the responses.
- If choices are letters, identify the alternatives with numerals, and vice versa.
- Avoid the options *all of the above* and *none of the above*.
- Avoid slang correct options.

Item Specifications:

Each multiple choice item should be written to specifications that can assure parallel item development as well as consistent item quality. Item specifications are “roadmaps” to developing similar items. An example of an item specification is shown below. The item is accompanied by the stimulus and response attributes. Teachers can write several items at the same level of difficulty and that assess similar math skills. The example was written by a team of teachers assigned to write math assessments suitable for placing students in a curriculum and monitoring their progress. Note that the specifications can be used to guide analysis during a cognitive lab session (see page 20).

<p style="text-align: center;">Problem 1:</p> $\begin{array}{r} 50,526 \\ -35,287 \\ \hline \end{array}$ <p>A. 15,239 B. 85,813 C. 15,339 D. 34,239</p>	<p>Stimulus Attributes</p> <ol style="list-style-type: none"> a. Subtraction problem written vertically b. Only base 10 whole numbers will be used c. Problem involves regrouping 4 times d. Subtracting ten thousands from ten thousands (the minuend has zeros in the tens and hundreds place) e. The minuend is larger than the subtrahend. f. Only one correct answer is larger than the subtrahend. g. Answer choices will be below. <p>Response Attributes</p> <ol style="list-style-type: none"> a. Four answers will be presented, one of which is accurate. Solution A is correct. b. Solution B is inaccurate because it is the sum of the two numbers and not the difference. c. Solution C is inaccurate because of borrowing errors in the hundreds place. d. Solution D is inaccurate because of borrowing error in the thousands place.
---	---

True-False

True-False items are perhaps the quickest to write, score and report but present challenges to reliability and validity. They have the advantage of assessing broad content which can mitigate some of the reliability and validity problems. True-False questions force a choice between only two possible responses and are generally used to test recall or comprehension. Some tips for writing true-false type items are:

- Target only one fact or idea at a time
- Avoid patterns of answers
- Make all statements about the same length
- Avoid absolute words like all, never, always, etc.
- Avoid indefinite adjectives like usually, generally, often, etc.
- Avoid complex sentences
- Use a connecting word like “because” when testing cause and effect logic
- Make false statements sound positive and avoid using negatives or double negative wording
- If negatives are used call attention to them by using italics, bold type, capital letters, or underlining
- Avoid using direct quotes from studied materials to discourage memorization

Table 1: Advantages and Disadvantages of Item Types

Type	Advantages	Disadvantages
Completion/ Short Answer	Reduces guessing. Can cover fairly wide content.	Limited range of abilities assessed. Limited machine scoring available. Must train scorers to assure uniformity.
Performance	Permits students to show work or proficiency.	Time consuming to prepare, administer and score.
Essay	Quick to construct. Eliminates guessing.	Restricts amount of content tested. Limited machine scoring. Must calibrate scoring and use anchors for inter-scorer reliability.
Matching	Easy to construct. Quick to score. Objective to score.	Generally used with lower level cognitive tasks.
Multiple-choice	Measure varying levels of student ability. Sample broad subject content. Quick and easy to score. Objective scoring. Open to robust statistical analysis.	Difficult to construct effective items. Must guard against measuring lower level cognitive skills.
True-False	Can test large sample of information. Quick to score.	Guessing. Difficult to construct effective items.

When determining the number of items to include, keep in mind that most formative assessments will need to be completed within a single setting of a typical class period. Guidelines in this area depend upon maturity level of students. In determining an assessment for high school level students the following guidelines are useful.

Table 2: Response Time Estimates by Item Type

Item Type	Average Time
True-False	30 seconds
Multiple choice	1 minute
Multiple choice of higher level learning objectives	1.5 minutes
Short answer	2 minutes
Completion	1 minute
Matching	30 seconds per response
Short Essay	10-15 minutes
Extended Essay	30 minutes
Performance	Varies

Designing a “Blueprint for an Assessment that is incorporated into Progress Monitoring using a Pacing Guide”

How a test is designed depends on the purpose(s) to be served. An instructional model that anticipates all students will master content at a given time will likely use few items but have them contain similar item difficulty. As stated in the Michigan State test writers guide: “Ideally, item discrimination (the degree to which an item differentiates between students with high test scores and students with low test scores) should be minimal in a mastery-model situation.” In a mastery-model we would like for all knowledgeable students to score high on items of similar difficulty.

Normative-model tests should have sufficient items across a spectrum of item difficulty that students will be spread according to their ability and content knowledge. More items are required to accomplish this purpose successfully.

Item difficulty and discrimination are not the same. It generally is easier to adjust item difficulty than item discrimination because discrimination relies on analysis within context of varying student ability. Difficulty is often a function of cognitive complexity. Cognitive complexity is guided by the mental operations required of the student to respond to the question. The key identifier of cognitive function is usually the verb incorporated into the question.

The following set of verbs is included to provide a quick reference when developing items of different levels of cognitive functioning. They are grouped according to the theory of cognitive ability developed by Benjamin Bloom and associates at the University of Chicago.

Table 3: Verb List

KNOWLEDGE	COMPREHENSION	APPLICATION	ANALYSIS	SYNTHESIS	EVALUATION
Define identify label list name recall recognize	collect comprehend describe discuss explain gather know locate observe paraphrase record restate review summarize tell understand	apply calculate choose demonstrate depict determine display estimate illustrate measure organize select show solve use	analyze ask categorize classify compare conclude conjecture contrast correlate differentiate distinguish edit examine explore group hypothesize infer interpret investigate predict relate research sort study	build combine compose construct continue convert create design develop expand extend formulate generalize integrate plan reason	assess critique debate evaluate judge justify revise

It is helpful to decide in advance the number of items that are intended to assess the different levels of cognitive functioning for which assessment is desired on the test. By defining the levels as part of the test design, subsequent tests can be constructed with a more parallel set of tasks for student groups to be tested with the alternate form.

Using a planning template such as the one illustrated below may be useful in test development. Cognitive labs (as described in the next section) can be useful for test validation.

Table 4: Item Allocation Planning Template

Cognitive Level	Standard & ELA code	Standard & ELA code	Standard & ELA code	Standard & ELA code	Standard & ELA code	Standard & ELA code
Knowledge	5 items					
Comprehension		5 items				
Application			5 items			
Analysis				5 items		
Synthesis					5 items	
Evaluation						5 items

Item Development and Validation: Using Cognitive Labs —“Thinking Out Loud”

Cognition is generally referred to as the “process of thought”. As students complete an assessment they process information to arrive at an answer to the question or solution to the problem. A cognitive lab employs a method of studying the mental processes used when completing tasks. This methodology grew out of a process first developed by Clayton Lewis while he was with IBM and later refined by Ericsson and Simon (1980, 1987, 1993). It has since been implemented in a variety of settings and is growing as a means of facilitating student learning. A rich discussion, complete with prompts and narrative of interactions between students and teachers can be found on the internet by doing a search on “Think Aloud Method” or “Cognitive Lab”.

We propose using cognitive labs as a routine part of developing local assessments and as part of teacher professional development for interpreting test results. When teachers ask students to think out loud regarding the mental process used to solve problems insights can be gained to inform instruction as well as strengthen test items where needed.

Teachers can modify think aloud strategies when teaching by interrupting instruction periodically to consider questions like:

- So far I’ve learned ...
- That was difficult to understand because ...
- That was interesting because ...
- I was confused by ...
- I wonder why ...
- The next thing to happen will probably be ...

Similar questions can be posed in math by considering questions like:

- If x is an odd number, then what is $3x$? Is it odd or even?
- Is $3x$ plus 1 odd or even?
- When solving this problem what must occur first?
- After you find the value within the parentheses what is the next step?

It is useful to ask rather general questions of students to gain insights into the cognitive dimensions of test taking that could guide formatting or item presentation features. These questions could include considering:

- Was this question easy, medium, or hard for you?
- Why did you do this first?
- Why did you stop?
- What did you like best (or least) about this item?

A blog posted by the Carnegie Foundation for the Advancement of Teaching noted a key reason that many students cannot solve complex math problems is that they do not have sufficient mastery of the underlying procedures required by the problem. The insight was uncovered by analyzing results from examining results of students thinking aloud to solve multiple step problems. The person who posted the blog postulated: “As a general rule, problems that require relevant, organized knowledge in long-term memory and a set of readily available routines that can be quickly searched during problem solving, presented extreme difficulties for the majority of the students. For many of these students, sub problems requiring simple arithmetic and algebraic routines such as the manipulation of fractions and exponents represented major, time-consuming digressions. In the vernacular of cognitive psychologists, the procedures were never routinized or automated. The net effect was that much solution time and in fact much of the students’ working memory was consumed in solving routine intermediate problems, so much so that they often lost track of where they were in the problem.”

Test publishers have used cognitive labs by employing think aloud methods successfully (Zucker et al. 2004). The method has also been effectively used to study test taking with such student groups such as English Learners or students with disabilities (Johnstone et al. 2006).

Advocates of the think aloud method recommend that the construct under consideration and solution strategies anticipated be identified prior to engaging students. By anticipating the types of errors to be encountered and the logic or problem strategy to be observed it is more likely that the strategy will produce useful information. Solving an algebra problem, for example, would likely produce very different response patterns than reading and interpreting a passage of narrative. There are some general considerations that apply across content areas. Reading level must be appropriate for the student, and opportunity to learn prior to testing are among the more obvious “rules” to observe.

When using think aloud methods it is important to approach students in respectful, non-threatening ways that permit each student to respond with minimum anxiety or sense of self doubt. The following example “think aloud” protocol script, cited in Technical Report 44 from the National Center on Educational Outcomes, clearly shows these principles.

Think Aloud Protocol Script

“We are interested in how students solve problems on tests, so we want to ask you and other students to solve some test problems for us and let us listen to how you do that. We are not as interested in the answer you come up with as we are with how you are thinking about the tasks.”

Notice the phrasing is general and honest about our interests and respectful of the contribution each student can make to tests for students across the country. Students should not feel the slightest sense of being judged or of having to obtain any particular types of results. Once they do, it affects their behavior and introduces a bias.

Ask the student to “parrot” back what he or she was told about today’s session by the recruiting person or teacher. Often, you will find that the student has been given information that is biasing and can affect the session. You need to find it in order to rectify it.

“What were you told we were going to do today?”

Be curious about what students do and why. Also tell the student that you will be videotaping the session and let him/her know when you turn on the camera.

“What you say is really important, so we are going to run this camera to make sure that we don’t forget anything.”

Provide practice.

Give each student a practice task to familiarize him or her with thinking aloud while working through a task. First you solve a problem and then ask the student to solve one. (The camera is not turned on for the practice.) Give the following instruction:

“I’m going to think out loud while I solve this problem. That means I’m going to say everything that goes through my mind.” (Complete problem while thinking out loud.)

“Now I’m going to ask you to solve a problem the same way. Just say everything that goes through your mind while you solve the problem.”

“I am not as interested in the answer to the problem as much as how you are thinking about the task. Do you have any questions about what we just did?”

When the think aloud process is coupled with sound analysis about the relationship between item difficulty and student ability the process should be strengthened. Teachers can better anticipate critical applications to interpreting reading or solving math problems when they have an alert about the complexity of the item presented or the proficiency of the student. Of course, care should be taken to not prejudge outcomes, and the teacher must remain objective as an observer of the process.

Acceptable and Unacceptable Questions

Examples of acceptable and unacceptable questions taken from released items that were used in state testing programs are included so you can practice applying what has been presented in the guide. Read the questions and consider why they may be acceptable or unacceptable for use in assessing student proficiency in the areas of mathematics and language arts.

EXAMPLE: ACCEPTABLE QUESTION – MATHEMATICS, 6TH GRADE

Skill description: This skill involves decimals in the form of currency and finding percentage discounts. All questions require students to determine the amount of a discount. No formulas are given; all numbers are less than 1,000 and decimals are no smaller than hundredths.

Jack wants to take Mary to the movies. He has a coupon for a 20% discount on two movie tickets. The price for one movie ticket is \$7.75. How much is Jack's discount on the two movie tickets?

- A. \$12.40
- B. \$6.20
- C. \$3.10
- D. \$1.55

This question clearly matches the skill description, has reasonable answer choices, grade level appropriate content, and contains no bias.

Answer: [C]

EXAMPLE: UNACCEPTABLE QUESTION – MATHEMATICS, 6TH GRADE

Skill description: This skill requires students to determine whether the problems require addition, subtraction, and/or multiplication. Some problems replace numerical digits (5) with word names (five). All numbers are less than 100, and decimals are no smaller than hundredths.

Sally's watch adds 5 minutes to every hour. She resets her watch every day at midnight.

When Sally's watch reads 6:00a.m., what time is it really?

- A. 3:45 p.m.
- B. 6:25 a.m.
- C. 5:35 p.m.
- D. 4:30 p.m.

Answer: [B]

This question is unacceptable because:

- It is not grade level appropriate.
- The "correct" answer is obvious because it is the only choice with a.m.
- This question doesn't completely match the skill description. It involves measurement.

EXAMPLE: ACCEPTABLE QUESTION – LANGUAGE ARTS, 4TH GRADE

Skill description: Students must identify and create simple sentences. Some questions will require students to convert fragments or compound sentences into simple sentences. In other questions, students must put the words in order to construct a simple sentence.

Which answer shows a simple sentence?

- A. Lowering myself to a crawl, I was able to creep beneath the house.
- B. My mother saw me outside by the house, and she yelled at me.
- C. I crept and crawled beside the house.
- D. Our house, in the middle of a street.

Answer: [C]

This question clearly matches the skill description, has reasonable answer choices, grade level appropriate content, and is written with the appropriate readability.

EXAMPLE: UNACCEPTABLE QUESTION – LANGUAGE ARTS, 5TH GRADE

Skill description: The learner will determine the correct combination of multiple sentences.

What is the best way to put these sentences together as one sentence?

The mortgage was too expensive for Luis to pay, so the bank was foreclosing on his house.

Paying bills on time was not one of Luis' strong points.

- A. Paying the majority of the bills on time was too expensive for Luis to pay, because the mortgage was foreclosing on the house so the strong point of Luis', the bank look at this closely.
- B. The mortgage was too expensive for Luis to pay, so the bank was foreclosing on his house. This is because paying bills on time was not one of Luis' strong points.
- C. The bank was foreclosing on Luis' house because the mortgage was too expensive to pay, and paying bills on time was not one of Luis' strong points.
- D. Too expensive was the mortgage, so the bank was foreclosing on his house, Luis', and paying bills on time was not one of Luis' strong points.

Answer: [C]

This question is unacceptable because:

- The question could be written more concisely: *What is the best way to combine these sentences?*
- Luis being unable to pay his mortgage and bills is biased.
- The question is inappropriate for the grade level because the concept of home ownership and paying bills is not something to which elementary or most high school students are exposed.
- It contains poor punctuation.
- The correct answer and foils are long.
- Answer choice 'B' still uses two sentences, so it is obviously incorrect.

WHERE TO FIND RELEASED ITEMS FROM STATE ASSESSMENT SYSTEMS

Most states annually release test items that have been used in statewide testing programs. Interested parties are directed to do a "Google" search on "released test questions". A website that provides links to most states released items is:

<http://www.edinfomatics.com/testing/testing.htm>

A few items are shown on the following pages. Often states release complete tests with directions for administration and actual student test booklets.

SAMPLES RELEASED FROM OTHER STATES

CONNECTICUT

Nick went to Dinosaur State Park in Rocky Hill and saw the fossilized dinosaur track shown in the scale diagram in your answer booklet.

Estimate the area of the dinosaur track using your centimeter ruler. Show your work or explain how you found your estimate.

Remember to show your work and write your answer in your answer booklet.

At a carnival booth, contestants pick a color on a large spinner. A prize is won if the arrow stops on the color they pick. The spinner is divided into 8 equal sections, as shown in your answer booklet. Each section is colored green, yellow, red, or blue.

The results for a sample of spins are shown in the chart below.

RESULT	# OF SPINS
Green	38
Yellow	58
Red	35
Blue	19

Use the results to predict the color of each of the sections on the spinner, and label each section of the spinner with the letter of the color: (G) green, (Y) yellow, (R) red, or (B) blue. Show the mathematics you used or explain how you decided how many sections should be labeled with each letter.

Remember to show your work and write your answer in your answer booklet.

CONNECTICUT

- 1** My Mom came to my room today and told me something that I've known for a while.
2 She said it was time to redecorate my room; it was long overdue, she said. I looked around
3 Uncomfortably and said she was right. It was time to face the decision I've been dreading.
4 When I was a kid, I was completely obsessed with dinosaurs. I read all about them and saw
5 every dinosaur movie ever made. I was an aauthority, able to rattle off any fact, no matter how
6 small, about any kind of dinosaur. I have gotten past this phase and my bedroom has not. I look
7 around and see dinosaurs everywhere, on the wallpaper, on the curtains, on the bedspread and
8 even hanging from the ceiling. The difficult decision is not whether to get rid of the dinosaur
9 décor, I know I have to do that. The question is, what should I use in place of it? It's a matter of
10 identity. I used to be a dinosaur kid. What kind of kid am I now?
11 I know I am not interested in dinosaurs anymore, but I have a problem. I don't have any
12 idea of what I want to do with this room. Should I choose something bold and dramatic? That
13 isn't really me. I could decide on something cool and subtle, but that isn't me, either. I'm not
14 artsy or retro or geometric or asymmetrical. I don't know how to match up my personality with a
15 decorating style. Does this mean I have to settle for a room that is totally beige? What kinds of
16 choices are there for an ordinary kid who used to love dinosaurs?
17 Before this develops into a full-blown crisis, I'll drive Mom to the shopping center. I'm sure
18 my mom and I will find something I like. We'll find some identity for me in a wallpaper book.

1. What is the **best** change, if any, to make in the sentence in **line 1** (***My . . . a while.***)?
 - a. Change ***Mom*** to ***mom***.
 - b. Insert a comma after ***today***.
 - c. Insert a semicolon after ***today***.
 - d. Make no change.
2. In the sentence in **lines 5-6** (***I . . . dinosaur.***), Jamie would like to change the word *small*. Which of these would be the **best** change for Jamie to make?
 - a. Common
 - b. Trivial
 - c. Scientific
 - d. Accurate
3. What is the **best** change, if any, to make in the sentence in **lines 5-6** (***I . . . dinosaur.***)?
 - a. Change ***dinosaur*** to ***Dinosaur***.
 - b. Change ***aauthority*** to ***authority***.
 - c. After ***aauthority***, change the comma to a semicolon.
 - d. Make no change.
4. What is the **best** way to change the sentence in **line 6** (***I . . . not.***)?
 - a. I have gotten past this phase, so my bedroom has not.
 - b. I have gotten past this phase, or my bedroom has not.
 - c. I have gotten past this phase, when my bedroom has not.
 - d. I have gotten past this phase, but my bedroom has not.

MASSACHUSETTS
English/Language Arts
Grade 4

WRITING PROMPT

Who is your favorite person to spend time with? Think of a special day or important time you shared with this person.

Think of a special time that you spent with your favorite person. Give enough details to show the reader what happened when you spent time with your favorite person.

The poem "The Photograph" is about a boy who watches his family study some photographs. Read to find out what happens to Mamá as she looks at photographs of her family and events of the past. As you read the poem, be sure to use the word bank to help you with the Spanish words and their meanings. Answer the questions that follow:

The Photograph

1 Mamá takes down
2 the large frame
3 with all of my cousins
4 my *tíos* and *tías*
5 and all of
6 the babies
7 the weddings
8 the birthdays
9 graduations
10 *quinceañeras*
11 *bailables*
12 *bautismos*:
13 Her little squares of México.

14 *Mamá* squeezes little pink Mimi
15 between my *tío* Ricardo
16 and the picture of her *quinceañera*.

17 *Mamá* was so beautiful then:
18 small shoulders inside her white dress,
19 her serious mouth,
20 her dancing eyes.

21 Mamá looks through
22 the glass
23 and the pictures
24 and the back of the frame
25 - clear through the wall.

26 She stands as still as her photograph.

Word Bank

Mamá – Mama
tío – uncle
tía – aunt
quinceañeras – special party for 15-year-old girls
bailables – dances with live music
bautismos – baptisms

MASSACHUSETTS
English/Language Arts
Grade 4

27 her eyes dance
28 like they did in her photograph.

29 She does not know
30 I saw her become
31 fifteen again.

- *Jane Medina*

"My Name is Jorge: On Both Sides of the River," text copyright ©1999 by Jane Medina. Published by Wordsong, Boyds Mills Press, Inc. Reprinted by permission.

1. What are the "little squares of México" referred to in line 13?
 - a. pages in an old photo album
 - b. pieces of pink material for clothes
 - c. places where people get together
 - d. photographs of family members

2. In line 27, what does "Her eyes dance" mean?
 - a. Her eyes move to music.
 - b. Her eyes appear gentle and wise.
 - c. Her eyes look excited and happy.
 - d. Her eyes fill with tears.

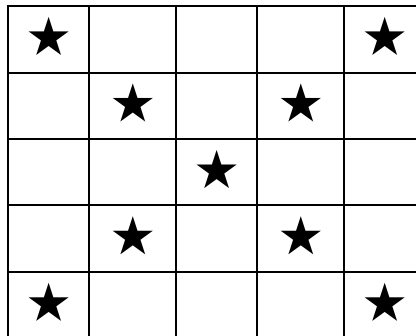
3. What is the **main** idea of lines 26-31?
 - a. The speaker begins to dance with *Mamá*.
 - b. *Mamá* finds an important photograph.
 - c. *Mamá* has special people in her life.
 - d. The speaker watches as *Mamá* changes.

4. Which of the following makes "The Photograph" a poem?
 - a. rhyming words
 - b. stanzas
 - c. stage directions
 - d. paragraphs

MASSACHUSETTS
Mathematics
Grade 4

1. Since $6 \times 3 = 18$, what is 600×3 ?
 - a. 180
 - b. 1,800
 - c. 18,000
 - d. 180,000

2. Yvonne used plain tiles and tiles with stars to make the design shown below.



- Which of the following fractions represents the part of the design that is made of tiles with stars?
- a. $\frac{1}{25}$
 - b. $\frac{1}{9}$
 - c. $\frac{9}{16}$
 - d. $\frac{9}{25}$
-
3. Lisa measured the length and width of the rectangular floor of her room. She used the measurements to find the area of the floor.

Which of the following could be the area of the floor of Lisa's room?
 - a. 120 square feet
 - b. 120 cubic feet
 - c. 120 inches
 - d. 120 yards

MASSACHUSETTS
English/Language Arts
Grade 10

WRITING PROMPT

Works of literature often feature characters with the ability to inspire or lead others.

From a work of literature you have read in or out of school, select a character with the ability to inspire or lead others. In a well-developed composition, identify the character, describe how the character inspires or leads others, and explain why this character's ability is significant to the meaning of the work of literature.

LANGUAGE AND LITERATURE

Imagine always seeing the letters of the alphabet in color or seeing shapes whenever you listen to music. This is the world some people experience. Find out more about this phenomenon by reading the Smithsonian magazine article "For Some, Pain Is Orange." Then answer the questions that follow.

FOR SOME, PAIN IS ORANGE

PERSONS WITH SYNESTHESIA EXPERIENCE "EXTRA" SENSATIONS.
THE LETTER 'T' MAY BE NAVY BLUE; A SOUND CAN TASTE LIKE PICKLES

BY SUSAN HORNIK

When New York artist Carol Steen was 7 and learning to read, she exclaimed to a classmate as they walked home from school, "Isn't A the prettiest pink you've ever seen?" Her little chum responded with a withering look. "You're weird," she said.

Shabana Tajwar was a bit older when she discovered that her world was more colorful than most. In 1991, as a 20-year-old intern, she and a group of friends were trying to remember someone's name over lunch. "I knew the name was green. It started with F and F is green," says Tajwar, now an environmental engineer. "But when I mentioned that, everyone said 'What are you talking about?'" She shrugs, "I was sort of in shock. I didn't know everyone didn't see things the same way."

While most of us experience the world through orderly, segregated senses, for some people two or more sensations are commingled.¹ For Steen and Tajwar, hearing a name or seeing a letter or word in black and white causes an involuntary sensation of color. To Tajwar the letter T is always navy blue. "I don't see the actual letter as colored," she says. "I see the color flash, sort of in my mind's eye." Steen not only delights in pink A's and gold Y's, she experiences colored taste as well. "I see the most brilliant blue after I eat a salty pretzel," she says.

Others with synesthesia – from the Greek *syn*, meaning together, and *aesthesis*, perception – may feel or taste sounds, or hear or taste shapes. The chords of a strumming guitar may be a soft brushing sensation at the back of an ankle, a musical note may taste like pickles, a trumpet may

MASSACHUSETTS

English/Language Arts

Grade 10

sound “pointed”, the taste of chicken may feel “round”. A teenager once confessed that her boyfriend’s kiss made her see “orange-sherbet foam”.

Even more baffling to outsiders: while synesthetes’ perceptions are consistent over time, they are not shared. Letters, for instance, don’t evoke the same color for everyone. Steen jokes that her good friend and fellow synesthete Patricia Duffy is “great” but misguided. “She thinks *L* is pale yellow, not black with blue highlights,” says Steen with a grin, as she pours a mug full of coffee in her downtown New York loft. Separately, over lunch in a sunny bistro, Duffy, a language instructor at the United Nations, confides “Some of Carol’s colors are so wrong!”

Even relatives who have synesthesia – it seems to run in families – see things differently. The Russian novelist Vladimir Nabokov tells in his memoirs about playing with a set of wooden blocks when he was 7 years old. He complained to his mother that the letters on the blocks weren’t the right colors. She was sympathetic. She, too, objected to the shades – though she also disagreed with some of her son’s color choices. According to one study, only one letter elicits consensus among a majority of synesthetes; apparently some 56 percent see *O* as a shade of white. For Nabokov, it radiated the hue of an “ivory-backed hand-mirror”.

People with synesthesia have described their unusual perceptions to intrigued but baffled researchers for more than 200 years. At times they were viewed as mentally defective, at other times idealized as artistically gifted. Often, they weren’t believed at all. Only in the past decade or so, using controlled studies, in-depth interviews and computer-aided visual tests, have scientists begun to identify

and catalog the staggering variety of these automatically induced sensations. “We’ve gone to great lengths to identify the range of forms,” says Peter Grossenbacher, a cognitive neuroscientist² and one of the foremost U.S. researchers on synesthesia. “We understand it’s a real experience. But we don’t know yet how it comes to pass.”

Already, scientists have discovered that synesthetes frequently have more than one form of the trait. Carol Steen’s tall-windowed loft – part living space, part art studio – is jammed with her synesthesia-inspired painting and sculptural models. Pulling letters painted on business-card-size pieces of paper off a shelf, she struggles to make clear the unique sensations that color her life and work. “It’s like viewing the world in multimedia,” she says. “I want to show other people what I’m seeing.”

What Steen is seeing is not only color triggered by certain sounds, smells and flavors’ when listening to music, she also sees shapes, which are reflected in her sculpture.

Steen also feels pain in color. When on vacation in British Columbia two years ago, she jumped down from a rock and tore a ligament. “All I saw was orange,” she says. “It was like wearing orange sunglasses.” In her paintings she depicts similar color sensations that she experiences during acupuncture. One abstract oil shows a green slash arcing through a field of red; in another a tiny red triangle drifts off into the distance on a sea of bright blue.

Researcher Peter Grossenbacher and a small cadre of scientists in this country, the United Kingdom, Canada, Germany and elsewhere are currently doing research with volunteers to try to figure out why Steen sees orange

MASSACHUSETTS
English/Language Arts
Grade 10

when the rest of us just ache. So far, they agree that synesthesia is more common in women than in men and is an international phenomenon. Grossenbacher primarily employs sophisticated screening and interviewing methods. Others, bolstered by dramatic advances in imaging techniques, are observing the neural activity of synesthetes and measuring the unique ways their brains respond to stimuli. In the process, they are shedding light on how we all perceive the world around us.

"It's the only way I know of perceiving," Steen points out. "If someone said they were going to take it away, it would be like saying they were going to cut off my leg." Although Steen delights in exploring her sensations, others remain ambivalent. When she was 20 and eating dinner with her family, Steen mentioned that the number 5 was yellow. "No," her father said. "It's yellow ocher."

¹ *commingled* – mixed together

² *cognitive neuroscientist* – a scientist who studies processes of the brain

"For Some, Pain Is Orange" by Susan Hornik, from *Smithsonian*, February 2001. Reprinted with permission of the author. All rights reserved.

1. How does the author use the title of the article?
 - a. to indicate that some people feel more pain than others do
 - b. to explain why some people like the color orange
 - c. to suggest new research about synesthesia
 - d. to attract the attention of readers who are unaware of synesthesia

2. The experiences reported in paragraphs 1 and 2 of the article **most likely** indicate which of the following?
 - a. Synesthetes tend to associate identical colors with the same letters.
 - b. Most synesthetes do not want to mention their unusual experiences to other people.
 - c. Synesthetes may not realize their experiences are unlike those of other people.
 - d. Most synesthetes experience synesthesia for the first time when they begin to learn letters.

-----NOTES-----

SECTION II: THREE FACETS OF ANALYZING FORMATIVE ASSESSMENTS

As a way of gauging individual and group progress, teachers regularly administer assessments to students in their classrooms. In order to address student misunderstandings of subject matter, it is important for teachers to know specifically what individual students know, what they can do with that knowledge, and what they do not know yet. Guidelines issued by professional organizations (e.g., National Research Council, 2001a), standards for teacher practice (AERA/APA/NCME, 1999; AFT/NCME/NEA, 1990), and research on the effects of classroom assessment on student learning (Black & Wiliam, 1998; Brookhart, 2004; Shepard, 2001; Wiliam, Lee, Harrison, & Black, 2004) document the importance of formative classroom assessment. While the goal is to use formative assessment to guide and improve learning, instead of just judging whether learning has occurred, results from past assessments can also help inform the design, interpretation, and use of future assessments.

Teachers typically design assessments, or choose commercially published assessments, to which they assign a weighted value toward the course grade. For example, teachers may make the first and second quiz in the unit worth 10 points each and the cumulative test at the end of the unit worth 50 points. This kind of assessment use allows teachers to measure student progress in a quantitative way. While a teacher may provide individualized feedback to students on each assessment, the feedback may not be tied to overall goals for learning in the unit. This guide looks at how the three facets of formative assessment can be used to help teachers interpret student work and learning outcomes.

Facet I focuses on analyzing single items in a test to identify students' misconceptions and consider instructional goals. The questions we attempt to answer with Facet I include:

1. What do attractive distractors in the most difficult items tell us about student misconceptions?
2. How are the most difficulty items reflected across standards?
3. How can I develop lesson plans to address student misconceptions?
4. How can I develop grade level instructional goals related to student performance?

Facet II focuses on analyzing groups of items in a test to identify commonalities across items to differentiate instruction. The questions we aim to answer with Facet II include:

1. What are some commonalities across the most difficult items that make the content so hard for students to master?
2. What are some commonalities across the easiest items that make them prerequisites for students to learn the content?
3. Looking deeper into the content of the items within each level, how might you describe the way students develop in their understanding of the content?

Facet III focuses on using appropriate scaling techniques and cut points to make informed programmatic decisions. Rather than arbitrarily making cut points to determine which students are in need of remediation, this approach allows the user to answer the following:

1. How can I identify students in my class that are struggling to meet proficiency on the CST?
2. How does my students' performance on the Benchmark help me predict actual performance on the CST?

This section walks through each of the Facets to demonstrate how locally developed assessments can be used formatively to inform classroom instruction, curricular mapping and programmatic intervention.

Facet I: Inform Classroom Instruction: Identify student misconceptions

When analyzing assessment results, teachers often rely on the student's overall score on the test, which doesn't provide enough evidence about students' particular misconceptions or provide diagnostic feedback to help students develop in their understanding of the content. Rather than focusing on the students' overall performance (e.g. Tristan answered 45% of the items correctly), we need to focus on performance of the items (e.g., 11% of the students answered item 1 correctly; shaded gray in Figure 1).

By focusing on item performance we can then conduct a structured item analysis to identify the items that are "most difficult" for students (i.e., the *items* with the lowest % correct) and look for attractive distractors that can pinpoint students' misconceptions associated with that particular content. Without item analysis, this level of detail to make important instructional decisions is missing.

Student	Item 1	Item 2	Item 3		Total % Correct
Samantha	A	A	B	80%
John	C	B	A	65%
Tristan	D	A	A	45%
		.			
		.			
		.			
Total % Correct	11%	84%	56%	

Figure 1: Item performance

To engage the participants in a data-driven dialogue to inform instruction (Lipton, & Wellman, 2004), we use a three-phase model.

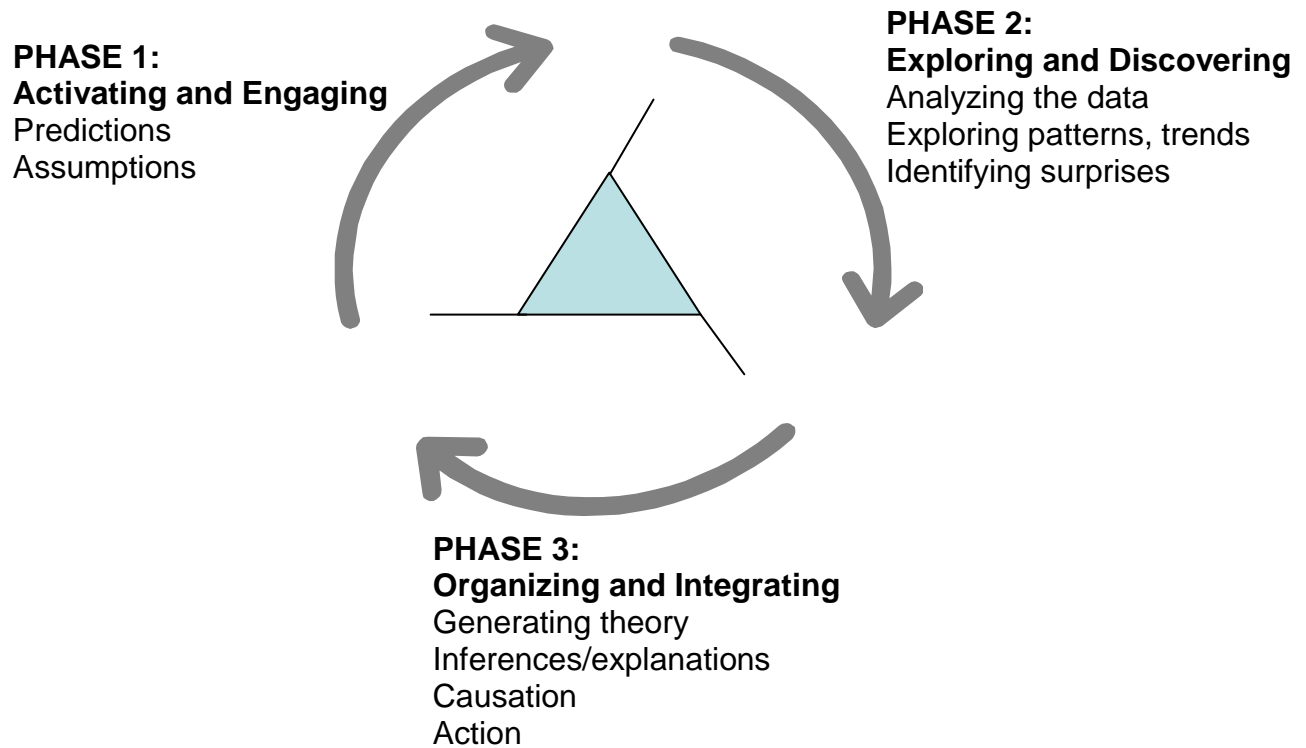


Figure 2: Three-Phase Model

Phase 1: Activating and Engaging - Making predictions and assumptions

We start by asking the participants to make predictions about item difficulty and to reveal their assumptions around why an item is harder or easier for the students. The trainer leading the discussion purposefully can select five items for participants to discuss, because they reveal misconceptions about students’ thought processes on particular content. A copied set of the items are cut and distributed to each group so that the participants can engage in a hands-on activity with the items.

Using the following graphic organizer participants identify the item choices that are likely to be attractive distracters for students, and write down their assumptions about why an item choice may identify a students’ misconception. The trainer may want to help the audience with a few sentence frames:

This item choice seems like a known misconception because _____.

This item choice may be attractive for my students because _____.

Predictions	Assumptions

Giving ample time for participants to discuss the answer choices of the items and the assumptions associated with these predictions is critical. This is when the participants are becoming actively engaged in the content of the test and aware of what to look for in the data once it is presented to them. For example, they may have discussed that item 35 is the most difficult item in the set of items presented, and have discussed that answer choice “B” makes the item more difficult, because they have seen this misconception among students in their class.

Phase 2: Exploring and Discovering - Analyzing the data for “attractive distractors”

Looking at the data from the item analysis, participants can affirm some of the predictions and assumptions they have discussed in Phase 1. Items should be sorted by difficulty from hardest to easiest, based on the percent of students who answered the item correctly (also known as a p-value) (Figure 3).

Answer Frequency								
Item No	Strand/Standard Aligned	Correct Response	Percent Correct	A	B	C	D	No Response
35	GR 08 Algebra I 15.0	D	11%	15%	57%	13%	11%	4%
32	GR 08 Algebra I 22.0	D	14%	26%	24%	31%	14%	5%
39	GR 08 Algebra II 22.0	B	14%	27%	14%	44%	9%	7%
14	GR 08 Algebra I 22.0	B	15%	47%	15%	20%	15%	3%
15	GR 08 Algebra I 6.0	D	16%	24%	19%	39%	16%	3%

Figure 3: Items sorted by difficulty (hardest to easiest)

Item 35 is the first item presented of the item analysis (Figure 3), since it is the “most difficult” item, with only 12% of students answering the item correctly. Since 59% of the students chose “B”, this answer choice (also called a “distractor”) was obviously the most “attractive”. The trainer might point out that answer choice “B” in item 35 as a good example of an attractive distractor, and then ask the participants to discuss if item 32 has an “attractive distractor”. The participants should recognize that the number of responses to the distractors is evenly distributed, so there is no “attractive distractor”. What makes a distractor “attractive” generally depends on the misconceptions and prior knowledge of students who responded to the questions.

For example, refer to Item 15 on the figure above. While 16% of the students got item 15 correct, 39% chose “C”. Therefore, “C” may be an attractive distractor. We must look deeper into the content of the item to identify the misconception that students have associated with item 15.

15. “A function has x-intercept 3 and y-intercept 2. Which of the functions below could be this function?”

A $4 + 3x = 2y$

B $2x - 3y = -6$

C $2y + 3x = 4$

D $3y - 6 = -2x$

Figure 4: Item 15

Item 15 aligns with a component within Algebra I Standard 6.0, in which “students graph a linear equation and compute the x- and y- intercepts (e.g., graph $2x + 6y = 4$). They are also able to sketch the region defined by linear inequality (e.g., they sketch the region defined by $2x + 6y < 4$).”

It is evident from the data that students who chose “C $2y + 3x = 4$ ” may have a misconception that may be related to their understanding of variables, in general, since they do not recognize x - and y -intercepts as points on a coordinate plane. These students may not know that they can substitute in values for the variables, x and y . If they understood this concept, they may have computed the y -intercept by substituting 0 for x and computed 2 for y , and then substituted 0 for y and computed $4/3$ for x (Figure 5). Instead of finding the equation that satisfies the two points (0, 2) and (3, 0), the students who chose “C” simply treat the 2 next to the y in the given equation as the y -intercept and 3 next to the x in the given equation as the x -intercept.

		A	B	C	D
		$4 + 3x = 2y$	$2x - 3y = -6$	$2y + 3x = 4$	$3y - 6 = -2x$
y-intercept is 2	X=0	$4+3(0)=2y$ (0, 3/2)	$2(0) - 3y = -6$ 6 (0, 2)	$2y + 3(0) = 4$ 4 (0,2)	$3y - 6 = -2(0)$ (0,2)
x-intercept is 3	Y=0	$4 + 3x = 0$ (-4/3, 0)	$2x - 3(0) = -6$ 6 (-3, 0)	$2(0) + 3x = 4$ 4 (4/3, 0)	$3(0) - 6 = -2x$ (3,0)

Figure 5: Tabular representation of function

Alternatively, or in addition, the students could have graphed the four lines with any values of x and y and found which of the four functions crosses the x -axis at (3,0) and the y -axis at (0,2). While some participants might debate if students may have rushed through the item and chosen “C” because at least one of the points satisfied the requirements, the trainer should emphasize the fact that the response for “B” is in standard form ($Ax + By = C$) would have been just as likely to be chosen if students were simply rushing through the test.

Phase 3: Organizing and Integration - Establishing next steps to undo misconceptions

To take participants to the third and final phase of Facet I, the trainer should hand out the “Next Steps” worksheet (Figure 7). This phase allows the participants to identify all of the student misconceptions represented in the data, and share best practices for undoing these misconceptions. The trainer should model one of the items for the participants. For example, using item 15 on the worksheet, the trainer would write the learning issue in the box: “Students don’t recognize that the intercepts are points on the coordinate graph with values for x and y ”.

To undo the misconception the trainer might suggest “Brain in the Hand” as a method for helping students become aware of their own misconceptions. In this activity, one student thinks aloud when solving a problem while another writes down the person’s thoughts. This particular example is also very conducive to helping students develop their academic language in mathematics. For example, a teacher might want to start students with a visual synectic and a sentence frame (Figure 6).



An intercept in mathematics is like an interception in football because _____.

Figure 6: Visual synectic and sentence frame

In sharing a student response, the trainer could share, “A student may consider contact with the football the same way that a line makes “contact” with the axes in a coordinate plane. A football player catches the ball at a particular point on the football field (e.g., 50-yard line on the right side) and then makes a path with his feet the same way a line makes a path across a coordinate graph.” Adding in a visual synectic and a sentence frame to help students discuss the mathematics may help students become more aware of the abstract concepts and develop their academic vocabulary that is so crucial to their success in mathematics. After sharing this example with the participants, the trainer should ask the participants to collaborate with one another and share their next steps with the entire group. With multiple participants in the training, this is generally a rare, but welcomed opportunity for discussion around cross-grade and within-in grade level articulation. The use of common assessments, such as district benchmark exams, provides an opportunity for participants to deepen their content knowledge, pedagogical skills, and use of data-driven instruction. By organizing the conversation around items on a common assessment, professional development may relate directly to pacing guide implications and instructional refinement.

Item #	% Correct	% Chose Attractive Distractor	Learning Issue - What misconceptions can you identify?	Teaching Issue - What can you do to undo these misconceptions?	Item/Test Issue - suggestions for revision	Next Steps -	
35	11%	57%					
39	14%	44%	<i>Participants fill in rest of worksheet in collaborative groups and share "next steps" with whole group.</i>				
14	15%	47%					
15	16%	39%	Students don't recognize that the intercepts are points on the coordinate graph with values for x and y	"Bird in Hand"; Visual Synectic (picture of interception in football)	Which of the "equations" below could be this function?	Find image of football player making interception. Choose some items for "Bird in Hand" activity	

Figure 7: "Next Steps" worksheet

Facet II: Inform Curricular Mapping: Recognize student’s development as a trajectory

The next level of analysis allows participants to look at the relationship between students and items on the same scale. Just as we oriented the participants to thinking about items from hardest to easiest in Facet I, we continue to discuss items in this order on a side-by-side map with students (Figure 8).

High Performing students	Harder items
Medium Performing students	Medium items
Low Performing Students	Easier items

Figure 8: Students and items on same scale

Compare students’ proficiency with item difficulty

In this graphic organizer, students are described generically by their performance on the left side of the map, and items are described generically by item difficulty on the right side of the map. As we consider how students progress in their understanding of the content, it is important to also consider what content can be used to measure that progress.

In order to deepen participants’ understanding of the data and to push participants to consider how students learn, from a cognitive perspective, we must select a measurement model that optimizes the interpretive quality of assessments. Rasch-based modeling (Rasch, 1961, 1980) provides a convenient way to develop estimates of student proficiency and item difficulty using the same scale. Mathematically, this model is represented as:

$$P_i(\theta) = P(X_i = 1 | \theta) = \frac{1}{1 + e^{-\theta - b_i}},$$

where, $P_i(\theta)$ denotes the probability of a correct response to item i and is solely a function of a student’s latent ability, θ , and the difficulty of the item, b_i . Based on probability of observed responses, the Rasch model allows us to analyze the developmental nature of the progress map, through a visual interpretive map, known as the *Wright Map* (Wright & Masters, 1982). The Wright Map, in conjunction with the progress map, provides a strong criterion-referenced interpretation of student proficiency. ConstructMap software, developed by Berkeley Evaluation & Assessment Research (BEAR) Center (Kennedy, Wilson & Draney, Tutuncuyan, & Vorp, 2006), is used for calibrating student ability and item difficulty.

Identify content in students' target "zone"

As demonstrated in the Wright Map for an Algebra I Benchmark Exam (Figure 9), students and items are placed on the same scale to consider mental operations and cognitive processes. For example, items 35, 34, 39, and 32 are the most difficult for students to master. Each X represents 18 students on the left hand side of the map. Therefore, there may be 18 students that have *actively learned* the content in items 35, 34, and 39. However, it is not assumed that they answered these questions correctly. We may say that these items are in their *target zone* (often referred to as ZPD, or Zone of Proximal Development, Vygotsky, 1978). These students distributed in the YELLOW have likely mastered the content represented by all of the items below their location on the map. Students distributed in the BLUE are the lowest proficiency on this test. They may be ready to learn the content represented in items 1 and 2. To keep things simple, we say that students distributed in the ORANGE:

- *have likely mastered* the content represented by the items in the BLUE,
- *actively learned* the content represented by the items in the ORANGE and
- *are ready to learn* the content represented by the items in the GREEN.

The items in the PINK and YELLOW may be too far from their target zone to focus on next during instruction. The distance between the student and the item determines the likelihood of answering the question correctly. The item is "within reach" if it is right next to the person. The item is said to be "out of reach" if it is far above the person. The item can be considered "too easy" if it is far below the person.

Since this particular test was not developed with an apriori theory, participants in the training should look for commonalities in items 35, 34, 39, and 32 to see what makes this content the most difficult for students to master. They may discuss language, cognitive load, a synthesis of ideas, complexity, etc. Additionally, they should look at the items at the bottom of the map to see if the content represents prerequisites for understanding the more challenging content. When time allows, participants can look deeper in the content of the items within the five colored bands to see if they can come up with a theory regarding how students develop in their understanding of the content (a backwards, neo-Piagetian approach to levels of how students develop).

Algebra I Math Winter 2008				
Scaled Score	Distribution of students (n=200)		Distribution of items (n=40)	
	High Proficiency ¹		More difficult items ²	
YELLOW				
	2			
				35
				34 39
			X	32
PINK	1			31
			XXXX	15 24 28 40
			XXXX	6 17 19 29 30 33 36
GREEN			XXX	16 20 21 26 27 37 38
	0		X	11 14 23
			-----XXX	
			XX	3 4 25
ORANGE			XXXXXXXXXXXXXXXXXXXXXXX	12 22
	-1		X	5 8 9 13 18
			X	
			X	7 10
BLUE			X	2
	-2		X	
				1
		Less Proficiency	Less difficult items	
		Each X represents 18 students	Cronbach's Alpha = .67	
		(---) Average Proficiency	Person Separation Reliability = .65	

Figure 9: Wright Map for Algebra I Math Winter 2008 Benchmark Exam

Since it is often a difficult task to hunt for a theory that may emerge from the data, we recommend using a confirmatory approach; that is, developing a test with an existing theory in mind.

¹ Student Proficiency = Calibrated person value which considers the difficulty of the items being answered correctly

² Item difficulty = Calibrated item value based on the percent of students who answered the item correctly

Validate and refine developmental model of student learning

The method used in this confirmatory approach applies the principles and building blocks of the Berkeley Evaluation & Assessment Research (BEAR) Center Assessment System (Wilson & Sloane, 2000; Wilson & Scalise, 2003; Wilson, 2005). The system is comprised of four building blocks, each associated with a core principle of the BEAR Assessment System. The principles ground the method at the intersection of learning theory with measurement theory. The building blocks include progress maps (also referred to as *progress variables* or *construct maps*), the items design, the outcome space, and the measurement model. Each building block is completed in an iterative fashion, always informing the next step, but often revealing desirable modifications to previous definitions.

Principle #1: Assessment should be based on a clearly defined developmental pathway for student learning. The building block to enact this principle is a set of one or more progress maps defining the “big ideas” in the curriculum for which you expect measurable development over time. Each progress map describes how knowledge in a particular domain develops over time.

Principle #2: What is assessed must be clearly aligned to what is taught—not the other way around. The building block for the alignment principle is the Items Design, which is focused on selecting just the right item content and format to assess growth on a particular progress variable.

Principle #3: Teachers are the principal managers and users of assessment data. The building block to implement this principle is the outcome space, which can be represented as a series of scoring guides, one for each item. An outcome space associates student responses with particular levels of knowledge on the progress maps. A scoring guide operationally defines the outcome space, and provides teachers with guidance for interpreting student work on particular items. If progress maps define the cognitive foundation of the assessment, then outcome spaces define the evidence base and the link to instruction.

Principle #4: To be most useful and fair, student assessment, whether formative or summative, must meet accepted standards of validity and reliability. The items developed to measure growth on progress maps should distribute themselves in accordance with the pathway set up in applying progress maps at the outset. In the BEAR Assessment System, the primary goal of selecting a measurement model is to optimize the interpretive quality of assessments. In order to provide a strong criterion-referenced interpretation of student proficiency, we place a priori interpretational constraints on the model during the design of items.

The Algebra I Progress Maps

As described above, a progress map describes a natural progression of knowledge, skills, or other competencies associated with the learning activities in a curriculum. It provides a common basis for interpretation across student responses on multiple tests and a common metric for measuring students over time. This building block is based on the idea that learning is developmental and may require students to overcome some conceptual hurdles. Developmental psychologists would agree that students must often conquer such conceptual hurdles to fully develop understanding in a particular area. Meyer & Land (2003) describe these hurdles as “threshold concepts”. When developing a curriculum, teaching an instructional unit, or administering an assessment to students, it is important to consider the developmental levels of students in the class. Teachers must consider misconceptions associated with the topic, as well as prerequisite knowledge that are necessary to fully understand the concepts. As such, the curriculum, instruction, and assessment must appropriately target students’ knowledge. Progress maps allow for effective interpretation of student learning and provide a basis for determining future instruction.

The aim of the progress map is consistent with the recommendations of the National Research Council (2001b). In their collaborative work, *Knowing What Students Know*, the NRC Committee describes the importance of thinking about student assessment on three critical, interacting aspects: Cognition, Interpretation, and Observation. Figure 10 shows the relationship among these three aspects. The bottom of the triangle, *Cognition*, can be viewed as the progress map, elaborating the cognitive model that is being measured. The left point of the triangle represents the *Observations*, the items that are designed to measure the construct. The right point of the triangle represents the *Interpretation*, the way in which the responses can be coded or scored so that they give information about the construct.

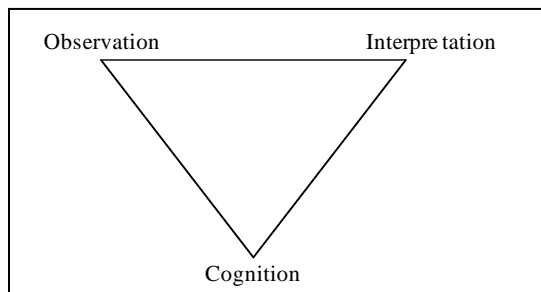


Figure 10. NRC Assessment Triangle (2001b)

The NRC recommendations have been further explicated in Wilson’s work on Constructing Measures (2005). He suggests that if we are going to try to measure a cognitive variable, we need to think of it on a continuum. The art of measuring depends on finding cognitive variables that are sufficiently simple to allow one to find an underlying continuum, but complex enough to be interesting. Rather than measuring students’ understanding as a binary trait (i.e. they understand it or they don’t), Wilson

(2005) suggests that some students may have more sophisticated understanding than others.

For example, when designing an instrument to measure Algebra competency, one might consider three possible continua for construct maps (Figure 11).

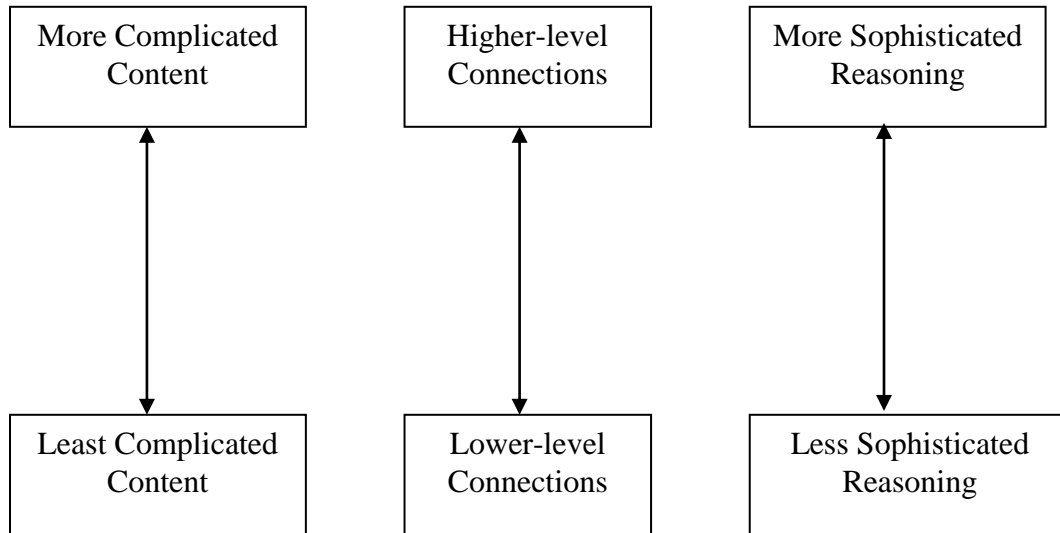


Figure 11: Possible Continua for Progress Maps

The first progress map describes a continuum to measure students' understanding of content. For example, in Algebra a student at the higher end of the progress map may understand more complicated functions, such as quadratic functions, while a student at the lower end may understand simple linear functions.

The second progress map describes a continuum to measure students' ability to make connections. For example, a student at the top of the progress map may have a deeper understanding of the content and be able to make multiple connections between graphs, symbols, and tables, while a student at the lower end may only be able to understand only parts of the content.

The third progress map describes a continuum to measure students' reasoning ability. For example, students at the top of the map may have more sophisticated reasoning ability when interpreting a graph — recognizing that the graph belongs to a family of functions (i.e. linear, quadratic, etc.). Students at the bottom of the map may lack algebraic skills and have difficulty in choosing appropriate points on the graph to interpret.

Thus, in developing the progress map, one should consider not only the domain that is being measured, but also how to adequately describe where the student is on the continuum of understanding. In order to create this developmental perspective about student learning in the form of a progress map, it is important to look at existing literature and to talk with experts in the field. Given the nature of the accountability movement and the practicalities around teachers' work in the classroom, it is also imperative to start the

development of a progress map by organizing state standards into a meaningful framework around big ideas.

For example, The College Board® developed formative assessments in their Springboard program (2006) to prepare students for success in college-level classes, including courses in the Advanced Placement Program in high school. Springboard is designed to offer rigorous content and uses the College Board Standards for College Success to “lay out a carefully articulated scope and sequence that builds knowledge and skills incrementally from sixth grade through twelfth grade” (Delgado, 2005). This existing structure from the CBSCS was used to design a progress map in the area of mathematical functions (Wilmot, 2008) Figure 12, below.

<i>Complexity of Functions</i>		
Level of Complexity	What the Student Knows	Response to items (repeats at every level)
6 - Trigonometric Polar Parametric	Student understands trigonometric, polar and parametric functions	Responses indicate that a student can:
5 – Exponential Logarithmic Recursive	Student understands exponential, logarithmic and recursive functions	-generalize this type of functions with a rule, - recognize/create/describe patterns from this type of function,
4 - Rational Radical Polynomial	Student understands rational, radical and polynomial functions	- create and extend patterns from this function with a rule,
3 – Absolute Value Piecewise Quadratic	Student understands absolute value, piecewise and quadratic functions	- create representations of this type of function,
2 - Multi-step Linear Inequalities	Student understands multi-step linear functions and inequalities.	- describe alternative representations of this function, - recognize/apply/translate among equivalent representations of this function,
1 – Simple Linear	Student understands simple linear functions	-compare/contrast equivalent representations of this function

Figure 12: “Complexity of Functions” progress map based on the Springboard’s CBSCS

This progress map, entitled the “Complexity of Functions,” is used to measure the learning trajectory of students’ college readiness as six developmental levels in the area of mathematical functions, and it is designed to offer a usable framework for teachers and professors to gauge student progress in this area of mathematical functions. The language is taken verbatim from the College Board’s Algebra content standard, “Patterns and Relations,” and the process standard, “Representations.”

In a College Readiness Assessment (CRA) developed by Wilmot (2008), fourteen multiple-choice items from College Board’s Springboard program were selected to map onto the six levels of the Complexity Construct Map. Three items map onto level one (L1A, L1B, L1C), three items map onto level two (L2A, L2B, L2C), three items map onto level three (L3A, L3B, L3C), two items map onto level four (L4A, L4B), two items map onto level five (L5A, L5B) and one item maps onto level six (L6A). These items represent a range of complexity in mathematical functions: simple linear, complex linear, quadratic, exponential, stepwise, and polar.

Because we have an existing theory, we can use a confirmatory approach to investigate the validity and reliability of the theory and the corresponding assessment items. Validity evidence, described below, is based on *internal structure*, *convergent evidence*, and *response processes*.

Validity Evidence based on Internal Structure

The Complexity Construct Map was designed according to the developmental learning progression specified in the College Board Standards for College Success for Integrated Mathematics (see Figure 4 in Chapter 3). By creating an intentional structure in the progress map, we can use a measurement model to analyze the fit of the items and to determine if the empirical results of the Wright Map (Wright & Masters, 1982) agree with the theory hypothesized in the Complexity Construct Map (Wilson, 2005).

To check the consistency and distinction of this progression, we can look at the Wright Map in Figure 13. This map shows a visual interpretation of the estimated student proficiencies (on the left side) and the estimated item difficulties (on the right side) after calibrating the items using the Rasch Model.

The Xs on the left hand side of the map represent the proficiency of 2356 students as distributed across the sample. There are fourteen items represented on the right side, with their respective levels from the Complexity of Functions progress map and the CBSCS written across the bottom of the map. The distribution of item difficulties covers the same region as the distribution of student proficiencies. That is, there are students represented at every level where there are items to measure them.

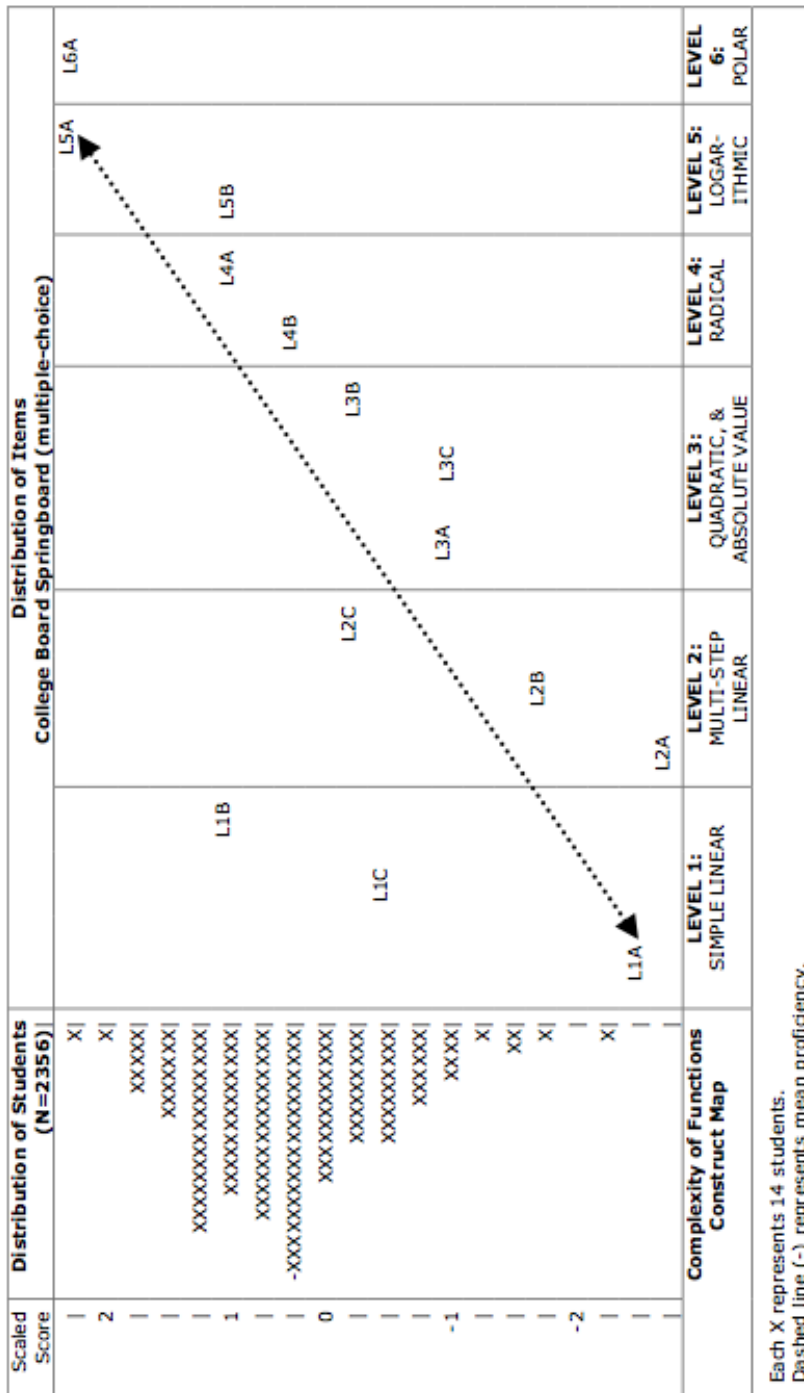


Figure 13: Wright Map for Complexity Construct Map

When looking at the distribution of item difficulties and student proficiency estimates on the Wright Map in Figure 13, we generally see a monotonically increasing trend from the easiest item to the hardest item (except for items L1C, L1B, L2C). As indicated by the dotted line in the Wright map, the items mostly appear to map onto the developmental progression stipulated in the CBSCS and the Complexity Construct Map. Items L1A, L2B, L3A, L3B, L4B, L4A, and L6A fall into quite close alignment with the developmental progression. Thus, with a quick glance at this item alignment across levels, it is clear that a cognitive framework, based on the CBSCS, may be adapted to consistently measure students' development in understanding the complexity of mathematical functions.

However, it appears that Level 5 items are interspersed with items in Level 4 and Level 6, and two of the Level 1 items are more difficult than expected, being located with items in Levels 3 and 4. In particular, the items that appear the most inconsistent with our expectations are from Levels 1 and 2 (L1C, L1B, and L2C). The item fit analysis and the verbal response data from teachers and students discussed in the sections *Convergent Evidence* and *Validity Evidence based on Response Processes* may offer some suggestions for this inconsistency.

Item fit analysis

Table 5 indicates the fit of the items. The first column is the name of the item. The second column is the calibrated item parameters. The last four columns include information about the fit statistics: the Form that shows the best fit, the infit meansquare, the t-values, and a judgment about the fit of the item³. As indicated in Table 5, only two items (L3B and L6A) may not fit well with the rest of the items on the test. Item L3B has a meansquare value slightly less than .75 and a t-value slightly less than -2, suggesting that student responses may have an overly regular response pattern. Sixty-eight percent (229 out of 338) of the students who answered L3B chose D, the correct answer. Thirty-one percent (105 out of 338) of students skipped the problem entirely, which may have resulted in this slightly poor fit compared to other items.

³ Infit meansquare is between 0.75 and 1.33, t-statistic between -2 and 2 (if one of these conditions is met, the item is considered to be a good fit) (Adams and Khoo, 1996, Wilson, 2005)

**Table 5: Item calibration estimates, and fit statistics
Springboard items**

Springboard Item	Calibrated Estimate	Fit statistics from Form	Infit Meansquare	t-value	Fits? Y/N
L1A	-2.44	E	.98	0.0	Y
L1B	1.00	F	.91	1.6	Y
L1C	-0.43	D	.75	-2.7	Y
L2A	-2.61	E	1.37	1.4	Y
L2B	-1.61	A	1.3	-2.4	Y
L2C	-0.12	D	.89	-1.2	Y
L3A	-0.95	G	1.25	1.4	Y
L3B	-0.05	D	.73	-2.3	N
L3C	-0.84	F	.83	-1.0	Y
L4A	0.79	E	.90	-1.6	Y
L4B	0.25	C	.90	-.5	Y
L5A	2.18	F	.91	-.5	Y
L5B	0.79	E	.95	-.7	Y
L6A	2.28	E	1.48	4.1	N

On the other hand, the meansquare value of 1.48 for item L6A suggests that student responses were more random than expected. As the most challenging item on the assessment, this item was obviously prone to lots of random guessing.

Convergent Evidence

In an attempt to compile convergent evidence regarding item difficulty, teachers were asked to rate the items on the test as easy, medium, and hard based on the students' mathematical experiences in their classroom. In many cases, the teachers suggested reasons why the items may be too easy or too hard. While this process may seem similar to the Angoff procedure (Angoff, 1971) which uses a person's judgment to identify cut points for a standard setting on a high stakes test, the purpose of teachers' evaluation in this research is not used to establish cut points. Rather, teachers' collective judgment, while it still may be unpredictable, is used to corroborate the difficulty of the items and to support the theory behind the calibration approach, which expects that students have varied educational experiences across the sample.

For example, one sixth-grade teacher explained in her pre-hoc evaluation of the items that many students will struggle with the term *linear relationship*, defined in an Algebra textbook as “a relationship that you can represent with a straight-line graph...characterized by a constant rate of change – that is, as the value of one variable changes by a constant amount, the value of the other variable also changes by a constant amount.” (Murdock, Kamishke, & Kamischke, 2002, p. 696). It is not surprising then that some of the College Board Springboard items (like L1B and L1C) were more difficult than expected, since the term *linear* is used in the item prompt.

Item L1B, a medium/hard item according to the calibrated item difficulty, asks students to choose the table where the relationship between x and y is linear (Figure 14). In a pre-hoc evaluation of the item, L1B was rated “hard” by teachers teaching sixth-grade math, seventh-grade math, pre-Algebra, “medium” by teachers teaching Algebra, and “easy” by teachers teaching Algebra II and pre-Calculus: this is approximately what one would expect.

L1B. In which of the following tables is the relationship between x and y a linear relationship?

A.	<table style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th style="border: 1px solid black; padding: 2px 5px;">x</th> <th style="border: 1px solid black; padding: 2px 5px;">y</th> </tr> </thead> <tbody> <tr><td style="border: 1px solid black; padding: 2px 5px;">1</td><td style="border: 1px solid black; padding: 2px 5px;">1</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">2</td><td style="border: 1px solid black; padding: 2px 5px;">4</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">3</td><td style="border: 1px solid black; padding: 2px 5px;">9</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">4</td><td style="border: 1px solid black; padding: 2px 5px;">16</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">5</td><td style="border: 1px solid black; padding: 2px 5px;">25</td></tr> </tbody> </table>	x	y	1	1	2	4	3	9	4	16	5	25	C.	<table style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th style="border: 1px solid black; padding: 2px 5px;">x</th> <th style="border: 1px solid black; padding: 2px 5px;">y</th> </tr> </thead> <tbody> <tr><td style="border: 1px solid black; padding: 2px 5px;">1</td><td style="border: 1px solid black; padding: 2px 5px;">2</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">2</td><td style="border: 1px solid black; padding: 2px 5px;">2</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">3</td><td style="border: 1px solid black; padding: 2px 5px;">4</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">4</td><td style="border: 1px solid black; padding: 2px 5px;">4</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">5</td><td style="border: 1px solid black; padding: 2px 5px;">6</td></tr> </tbody> </table>	x	y	1	2	2	2	3	4	4	4	5	6
x	y																										
1	1																										
2	4																										
3	9																										
4	16																										
5	25																										
x	y																										
1	2																										
2	2																										
3	4																										
4	4																										
5	6																										
B.	<table style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th style="border: 1px solid black; padding: 2px 5px;">x</th> <th style="border: 1px solid black; padding: 2px 5px;">y</th> </tr> </thead> <tbody> <tr><td style="border: 1px solid black; padding: 2px 5px;">1</td><td style="border: 1px solid black; padding: 2px 5px;">3</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">2</td><td style="border: 1px solid black; padding: 2px 5px;">6</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">3</td><td style="border: 1px solid black; padding: 2px 5px;">9</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">4</td><td style="border: 1px solid black; padding: 2px 5px;">14</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">5</td><td style="border: 1px solid black; padding: 2px 5px;">19</td></tr> </tbody> </table>	x	y	1	3	2	6	3	9	4	14	5	19	D.	<table style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th style="border: 1px solid black; padding: 2px 5px;">x</th> <th style="border: 1px solid black; padding: 2px 5px;">y</th> </tr> </thead> <tbody> <tr><td style="border: 1px solid black; padding: 2px 5px;">1</td><td style="border: 1px solid black; padding: 2px 5px;">5</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">2</td><td style="border: 1px solid black; padding: 2px 5px;">7</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">3</td><td style="border: 1px solid black; padding: 2px 5px;">9</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">4</td><td style="border: 1px solid black; padding: 2px 5px;">11</td></tr> <tr><td style="border: 1px solid black; padding: 2px 5px;">5</td><td style="border: 1px solid black; padding: 2px 5px;">13</td></tr> </tbody> </table>	x	y	1	5	2	7	3	9	4	11	5	13
x	y																										
1	3																										
2	6																										
3	9																										
4	14																										
5	19																										
x	y																										
1	5																										
2	7																										
3	9																										
4	11																										
5	13																										

Figure 14: Springboard Item L1B

Item L1C, a medium item according to the calibrated item, asks students to identify the graph that shows a linear relationship (see Figure 15). During pre-hoc evaluations, teachers teaching sixth-grade math, seventh-grade math, and pre-algebra all rated this item as “medium”, while teachers teaching Algebra I and above (up through pre-calculus) rated this item as “easy”.

L1C. Which of the following graphs shows a linear relationship between x and y ?

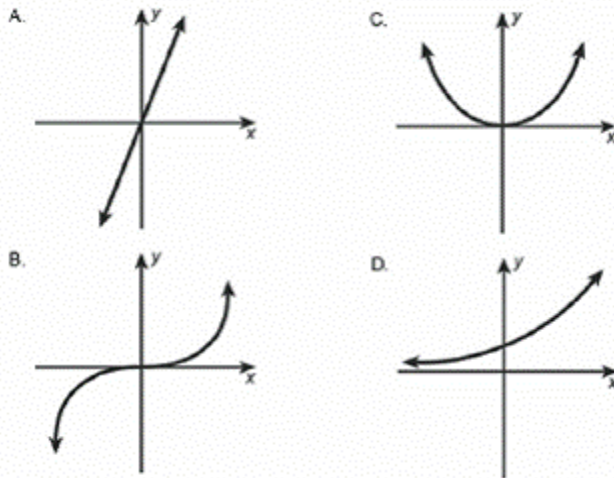


Figure 15: Springboard Item L1C

Item L1A, an easy item according to the calibrated item difficulty, includes a table of values for x and y , and asks students to identify which equation represents the linear pattern in the table (Figure 16). Math teachers teaching sixth grade math, pre-algebra, and geometry all rated this item as easy.

L1A.

x	y
1	5
2	9
3	13
4	17
5	21

Which of the following equations represents the linear pattern shown in the table above?

- A. $y = x + 4$
- B. $y = 4x + 1$
- C. $y = 2x + 3$
- D. $y = 3x + 2$

Figure 16: Springboard Item L1A

Although there are many examples of this type of agreement between item difficulty and the remarks of teachers about the content of the test, one last example is presented below. Item L5A, a hard item according to the calibrated item difficulty, asks students to identify the type of function represented by the table of values (Figure 17).

Three Algebra II teachers thought that L5A would be hard for students. For example, one Algebra II teacher said, “They won’t know what logarithmic means and they will be confused about the difference between exponential and quadratic [functions].” Another Algebra II teacher said that they “didn’t cover [the topic of logarithmic functions] yet”...they “have only worked with linear [functions] and a little bit of quadratic [functions] at this point in the year”. One geometry teacher said that the “vocabulary [in L5A] may be difficult for many [students].” According to one pre-calculus teacher, Item 5A would be “easy as long as [students] are familiar with the different types of functions”.

These difficulty ratings for the items suggest the learning opportunities students have encountered before taking the test. Therefore, this type of agreement provides additional evidence towards the soundness of the instrument.

L5A.

x	y
4	1
16	2
64	3
256	4
...	...

The table above gives some of the function values for $y = f(x)$. The domain of the function f is $x > 0$. Which of the following could describe the relationship shown in the table?

- A. exponential
- B. logarithmic
- C. linear
- D. quadratic

Figure 17: Springboard Item L5A

Validity Evidence based on response processes

Validity evidence based on response processes, which is based on students' interpretations of the assessment items, was collected for the multiple-choice items after students finished taking the CRA. At least three students in each class (one for each form) were randomly selected to participate in an *exit interview*, a quick one-to-two minute conversation where students were asked to identify any of the questions that were too easy or too challenging or any language they didn't understand.

For example, ten students from eight different classrooms (sixth-grade math through pre-calculus) said that L1A was one of the easiest items on the test. Seventeen sixth-grade students across eight classrooms mentioned that they didn't understand the term *linear*. The confusion around the definition of *linear* was evident in exit interviews with students in seventh-grade math, Algebra I, Integrated Mathematics I, Geometry, and Discrete math. One pre-calculus student said that item L5A (calibrated as the second hardest multiple-choice item on the test) was difficult because he "didn't understand the term logarithmic". These student exit interviews suggest that the calibrated item difficulties may be accurate, and offer additional evidence towards the validity of the instrument.

The distribution of the calibrated item parameters suggests four levels of complexity, rather than six. Level 1 and Level 2 items appear to group together and Level 5 items are split between Level 4 and Level 6. The findings from this study recommend a revision to the Complexity Construct Map. That is, four levels of complexity: Level 1 - All Linear, Level 2 - Quadratic & Absolute Value, Level 3 - Radical & Exponential, Level 4 - Polar & Logarithmic. Since there are only fourteen Springboard items represented on this test, it is difficult to know if these findings are a result of the particular items selected or if this is representative of items across the Springboard program. These results may be further substantiated with additional research with more Springboard items.

Discussion

It is not entirely surprising that all of the Level 1 items turned out to be of different empirical difficulty levels. Although one would expect that all of the Level 1 College Board Springboard items to be around the same difficulty, item L1A is the only item that aligns well with the developmental progression stipulated in the CBSCS and the Complexity Construct Map. By looking back at the Wright Map in Figure 13 it appears that Item L1C is as difficult as the Level 3 items included on the test, and item L1B is as difficult as the Level 4 items included on the test. This could be a result of the true empirical difficulty of the items, or it could be a result of a curricular mismatch.

According to the CBSCS, these Level 1 items measure students' ability to "recognize/apply/translate among equivalent representations" of "simple linear" functions. In short, item L1A asks students to translate among the verbal representation (*linear*), the symbolic representation (*equation*), and the tabular representation of a simple linear function; item L1B asks students to translate among tabular and verbal representations of a simple linear function; and item L1C asks students to translate among the verbal representation and the graphical representation (*the straight line*).

The exit interviews with students and the pre-hoc teacher evaluations of the items suggest that many students struggled to correctly answer items L1B and L1C because they did not understand the term *linear*. This was especially the case for the students in Sixth-grade math. This distinction can be seen even more clearly in Table 6 (following). For example, when we compare the percent of students who got L1A correct (which represented a Level 1 question as expected) with the percent of students who got item L1C correct, the striking difference is that 76% of students in middle school math got L1A correct, but only 45% of them got L1C correct.

Table 6: Percent correct across Level 1 Springboard items by grade level

Items at Level 1	Percent Correct on Item		
	L1A – T to S	L1C – V to T	L1B – V to G
Middle School: 6 th grade Math, 7 th grade Math, Pre-Algebra, Algebraic Concepts	76%	45%	27%
Lower Division High School: Algebra I ⁴ , Geometry, Double Block Algebra, Integrated Math	72% ⁵	75%	34%
Upper Division High School: Algebra II, Trig Honors, Pre-Calculus, AP Statistics, Calculus AB/BC, Functions/Statistics/Trigonometry, Discrete Math	87% ⁶	85%	59%

Why was L1A so much easier for students? Item L1A includes a table of values, the term *linear* in the prompt, and four equations. One might argue that students do not need to know what *linear* means (i.e., the verbal representation) in order to solve item L1A. This is because students only need to plug in values for x to find y and identify which equation works for all of the values in the table. In fact, one teacher commented that the “definition of linear would be a problem” in item L1A “but [students] can plug in the numbers [to find the correct solution]”. Thus, to correctly answer item L1A, students don’t really need to know that the correct equation is characterized as a linear function or why the coefficient of x (i.e. the slope) is a “4” and the y -intercept is a “1”.

Thus, the connection that students are making between representations in this item is not clear. Perhaps this item is easy because it is probably measuring students’ ability to substitute values for variables and do arithmetic correctly, which is a necessary prerequisite to understanding linear functions.

On the other hand, item L1C may be more difficult than expected because students must recognize that the term *linear relationship* is applicable from a picture of the graph. The exit interviews with the middle school students suggested that they were confused by the term *linear*. In some cases, students even pronounced it as “lye-nee-er”. So, while some of these students could deduce that the term *linear* was a derivation of *line* and pick the correct graph, other students simply could not make the connection.

It appears that Item L1B is even more difficult because it asks students to recognize which table of values is *linear*. Students cannot use their intuition of what linear might

⁴ This includes 8th grade Algebra I students as well.

⁵ This is the percent correct from Form C. Form D students performed a little bit better (51% got correct).

⁶ This includes students across the sample, since it was on Form E, the calibration form.

look like in a graph. That is, they cannot rely on their everyday experience of drawing lines in school to make a connection with the graph, as they could in L1C. Instead, students must grasp the concept of the ordered pair, a string of inputs and outputs, and recognize a constant difference in the y values (since the x -values are well-ordered). Even in upper division high school math classes (i.e. Algebra II and above), 41% of students got this item incorrect.

While we can expect the middle school students to get L1B wrong, since they were generally unfamiliar with the term *linear*, it is quite surprising that students in upper division high school math classes struggled to identify the table of values that represent a linear relationship. Perhaps this is a reflection of the manner in which linear functions are taught in the schools (predominantly through equations and graphs like Item L1A).

Based on the results discussed above, one might conclude that most students in sixth through twelfth grade struggle to make connections between the verbal and tabular representation of linear functions. And, it is clear that middle school students who are not yet in Algebra I are struggling to recognize graphical representations of linear functions. Thus, one might conjecture that it might not only be the *type* of function that makes these items easy or hard for students, but also the *kind of connections* between representations that students are expected to make.

It is also quite possible that this finding stems from a curricular issue, because students may not have had the same kind of exposure to the material in this way. Students probably have an easier time graphing symbolic expressions and plotting the data in tables, but have trouble “seeing” the tabular presentation of a linear function as being linear because they haven’t been asked to do so. Were the curriculum arranged differently, with those connections being explicit, students would likely have a better chance at making those connections. This finding is corroborated and discussed in more detail in Wilmot (2008).

Facet III: Inform Programmatic Intervention: Understand students’ needs

While some of the discussion above centered on compiling evidence for validity and reliability of formative assessments, this book does not exhaust the opportunities for test developers to evaluate their own assessments. There are many different types of evidence for determining validity and reliability. Reliability evidence may include internal consistency indicators, such as Cronbach’s Alpha (Cronbach, 1990), person separation reliability (Wright & Masters, 1982), and inter-rater reliability. Validity may be investigated by looking for evidence based on internal structure; that is, the proposed levels of the cognitive theories in the progress maps compared to the empirical levels suggested by the Wright Maps, and a detailed analysis of the items. In addition to describing the psychometric properties, detailed examples of student work, and feedback from teachers may report validity evidence based on instrument content, and response processes (Wilson, 2005).

For benchmark assessments, specifically, validity may also rest on the predictive validity of the assessment to the California Standards Test (CST).

Unfortunately, some districts mistakenly create cut points that are unnecessarily high. Using the A-F model of evaluation they tend to assign raw scores of 90% and above to Proficient performance level, and 50% and below to the Far Below Basic performance level (Figure 18).

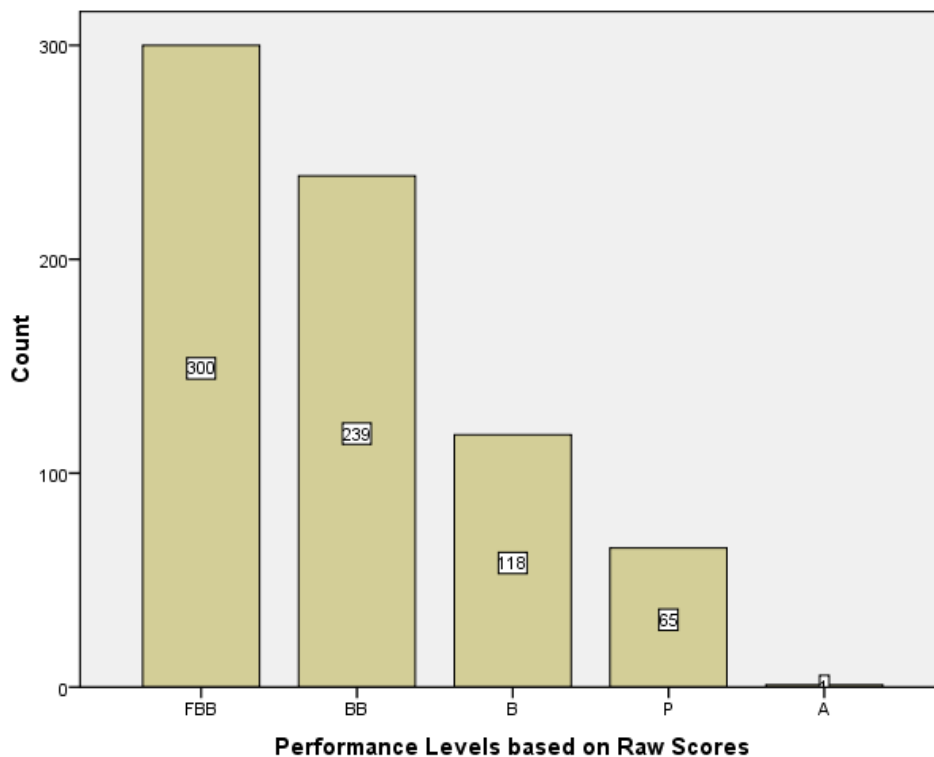


Figure 18: Performance levels based on Raw Scores

When matched up to actual scaled scores and their associated cut points (Figure X), it is easy to see how the predictive validity of these tests would be in jeopardy. Not only are the tests' validity in question, but the assignment of students into intervention programs may also be misguided as a result of inaccurate cut points based on raw scores.

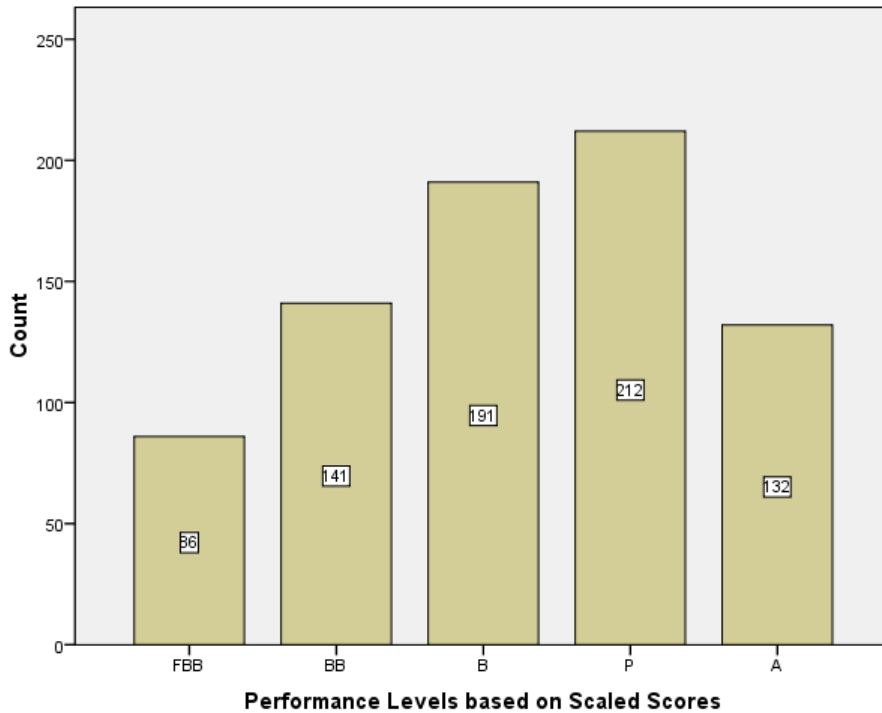


Figure 19: Performance levels based on Scaled Scores

Moving beyond a conjecture that students need remediation

Facet III addresses this predictive validity by assigning a scaled score for each student that is comprised on the CST metric. By predicting performance on the CST with a scaled score, districts can accurately monitor student progress towards proficiency on the CST and answer the following questions:

1. How can I identify students in my class that are struggling to meet proficiency on the CST?
2. How does student performance on the Benchmark assessment reflect their predicted performance on the CST?

Moulton (2007) has developed EDS-scaled scores for locally developed benchmark assessments to help districts answer these very questions. His approach takes the guessing out of cut-points and proficiency levels and uses a sophisticated mathematical model to predict CST performance today, and on the day of the actual test. For example, there are several students in the 3rd grade who got a 60% on the Benchmark Test in Mathematics. However, not all of them will get the same scaled score. As shown between the dashed lines (Figure 20), the student represented by the circle at the top has an EDS-scale score near 500, which would put her in the Advanced performance level, while the students near the bottom of the 60% group are scoring near a 300, which is indicative of the Basic level. These are critical pieces of information to help educators make smart decisions about student placement, master scheduling, and curriculum pacing.

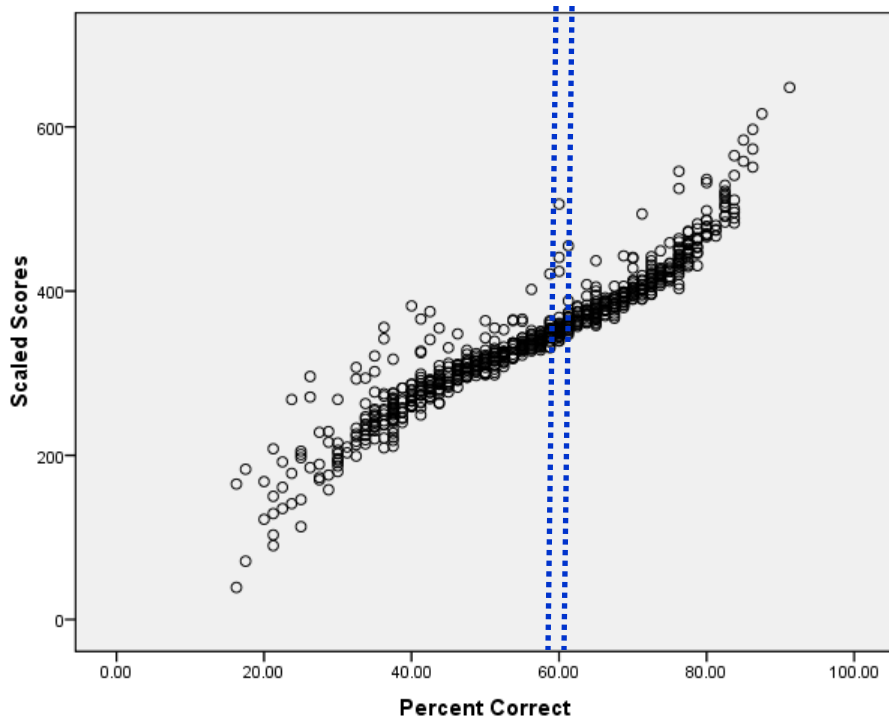


Figure 20: Relationship between EDS-Scaled Scores and Percent Correct

With scaled scores directly connected to the CST and their corresponding cut points for performance levels, teachers, principals, and district administrators can monitor progress towards meeting schoolwide goals. While it may be true that this idea of predictive validity was not the original intention of formative assessment, there is a critical need for school officials to have accurate data in this era of accountability. Moulton’s innovative approach (described in more detail in Section III) offers teachers and school administrators both accurate data and peace of mind that programmatic decisions can be based around students’ real learning needs.

-----NOTES-----

SECTION III: USING FORMATIVE ASSESSMENTS TO PREDICT PERFORMANCE

How Benchmark Exams Can Be Turned into Mini-CSTs

Why Local Benchmark Exams?

Local benchmark exams can shed light on students' strengths and weaknesses. The data that is closest to instruction should be used to inform teaching. In other words, won't we be helping students and teachers even more if we practice assessment "for learning" in addition to assessment "of learning"?

Is there data that simultaneously guides teachers and parents in understanding the academic growth of their students and is sufficiently rigorous, valid, and reliable that it can serve as an accountability indicator of school performance?

An answer to that policy question can be found in educational practice on a broad scale. The answer uses a similar methodology to one presently used in the STAR program's calculations of Lexile reading levels for students; it is a latent trait application of vertically scaled academic achievement data.

For the application to be feasible it must conform to a set of criteria that does not risk throwing out the work that has been accomplished in implementing the California curriculum standards. Some of these criteria are:

- Must be based on State standards
- Must be applied in a formative context
- Must be compatible with the present summative assessments of the STAR CSTs
- Must comply with NCLB requirement of assuring that students make AYP
- Must support program evaluation by measuring student level growth across grades.

Local benchmark exams, as now used, fail to meet these criteria.

Difficulties with Local Benchmark Exams

Formative assessments in the form of "benchmark" exams administered district-wide two-to-six times per year have become widely used in California districts in the wake of NCLB. Developed by the district or purchased from an educational vendor, they provide – or are intended to provide – guidance to district administrators, school principals, and teachers regarding several important questions:

- Are students districtwide "on-track" to score "Proficient" on the end-of-year CSTs?
- Are students districtwide meeting benchmark standards defined by the district?
- How are students, classrooms, and schools performing relative to each other at a given moment in time?

- What schools, programs, and other forms of educational implementation are proving most effective and least effective in improving student achievement over time?
- What content standards are students having the most trouble with — districtwide, schoolwide, and at the classroom level?
- What are the strengths and weakness of individual students?
- How quickly are students growing academically?

While benchmark exams gather a substantial amount of individual student-level information, districts often find it difficult to use their benchmark test results to answer the questions for which they originally purchased the exam. Benchmark exams are not equated to the CSTs, so it is problematic to infer a likely “percent Proficient” statistic from them. Benchmark standards vary from test to test and use cut-points that have often not been decided using a rigorous standard-setting procedure.

Attempts to measure school, program, and classroom effectiveness require some type of gain-score to capture growth over a period of time. Benchmark exams are unable to measure growth since they are not equated to each other. Indeed, they lack even a clearly defined construct in terms of which to measure growth. Each exam assesses performance in its own unique content in a raw “percent correct” metric and is not written to be comparable to any other exam. This makes valid program evaluation impossible using benchmark scores.

Attempts to diagnose strengths and weaknesses are similarly hamstrung because there is no effort to control item difficulty. Item p-values vary for many reasons that have little to do with student competence in the subject area; mere similarity between a distractor and the correct answer can convert an easy item into a very difficult one. When students score low on a content standard, it is hard to decide whether this indicates a legitimate weakness in the examinees or merely the presence of a set of items that are difficult for technical reasons. At the level of the individual student, there tends to be too few items per content standard to allow valid measurements of that student on that standard.

The only question that benchmark exams *can* answer successfully is: How are students, classrooms, and schools performing relative to each other at a given moment in time? Therefore, it appears that if districts are to obtain useful answers from their benchmark exams, alternative methods of analysis and scaling must be used. By applying multidimensional equating methods to local benchmark exams, a procedure is proposed that addresses these and related issues by converting locally developed exams into “mini-CSTs.” This procedure was developed by Educational Data Systems (EDS) in collaboration with the Santa Clara County Office of Education specifically to answer the questions and meet the criteria cited above.

The Benchmark Scaling Method Used by Educational Data Systems

The problem with benchmark exams is that they are not equated, either among themselves or in relation to the CSTs. Equating tests over time requires two conditions:

- A common construct, so that all benchmark exams measure along the same dimensions
- Common items, to compute the relative difficulties of benchmark exams

Unfortunately, neither condition is met with benchmark exams. In order to scale benchmarks using exam data as it currently resides in district databases (i.e., without requiring districts to administer new tests to equate existing ones), it is necessary to compensate for these two deficiencies.

Dimensional Alignment

Before calculating the relative difficulty of benchmark exams, it is necessary to realign each benchmark exam to a common construct or dimension. That means deciding on a construct.

One construct that naturally lends itself is that defined by the CSTs for reading and mathematics. Unlike benchmark exams, the CSTs are scaled specifically to embody a definite construct according to the measurement demands of the psychometric model that Educational Testing Service used to scale them. While in theory these constructs may vary from grade to grade according to the content standards for each grade, in practice we find through factor analysis that reading and mathematics embody reasonably coherent dimensions that extend across adjacent grade spans, allowing for the *possibility* of a vertically articulated cross-grade common scale analogous to that recently introduced into the CELDT exam. This should not be surprising given the use of cross-grade reading and mathematics vertical scales in other states and by such organizations as the Northwest Evaluation Association, which use similar types of items.

To perform the alignment, EDS uses a multidimensional IRT algorithm called NOUS (Moulton, 2005). Item level benchmark data is merged with the district STAR file from which are selected the reading and math CST scale scores that each student received at the end of the previous school year. After preparatory analysis to convert each student response (including choice of distractor) into a standardized metric, NOUS is applied to only the benchmark data to locate each student in a 2-dimensional space which is assumed to span the dimensionality of the corresponding CST exam for the previous year. The 2-dimensional solution was chosen as a default since it has been found to be optimal, or close to optimal, for most benchmark exams given the richness of the distractor-level response data.

Once students are located in the 2-dimensional benchmark space, their spatial coordinates are anchored and their CST scale scores, suitably standardized, are introduced into the data set. NOUS is now applied to calculate the two spatial coordinates of the CST variable. Because NOUS has erected a space anchored to the 2-dimensional coordinate

space calculated for the students based solely on their benchmark scores, the CST variable has in effect been *projected* into that 2-dimensional subspace. All sources of variance in the CST scores are *not* explained by the benchmark scores – in particular the different growth rates of the students since they took the exam the previous spring — are automatically filtered out. The 2-dimensional person vectors are matrix multiplied by the 2-dimensional CST vector to calculate an expected CST score for each person based on their performance on the benchmark exam. Having filtered out the effect of time (the different student growth rates since the previous spring), and if our assumption is correct that the CST exam spans the same 2-dimensional space as the benchmark exam, then we have in effect answered the question, “*What **would** each student have scored on the CST (on a non-difficulty-adjusted standardized metric) had he or she taken it at the same time as the benchmark exam?*”

If our assumption is false and the CST in fact covers content not statistically present in any form on the benchmark exam, then the expected CST score computed by NOUS will be restricted to only that aspect of the CST that is covered by the benchmark exam, manifesting as expected CST measurement error and a corresponding departure from the common scale. If the benchmark exam does erect a space that includes the CST exam as a subspace, but adds content that is not statistically present on the CST, the expected CST scores will not be affected.

That is the procedure by which EDS mathematically aligns each benchmark exam to a common CST construct.

Adjusting for Difficulty

While the dimensional alignment process aligns the benchmark exam with the CST dimensionally, it does not adjust for the fact that it has a different difficulty than the CST that was administered in the previous spring. Nor does it adjust for the fact that the individual benchmark exams have different difficulties, though it is reasonable to suppose that their difficulties are likely to increase through the school year to keep pace with student growth.

Without common items there is no direct way to compare the relative difficulty of two benchmark exams, or (an alternative way of saying the same thing) to compare the relative average abilities of the students who take the two exams. What *is* known is when each benchmark exam was administered. Therefore, EDS worked out a process for measuring the relative difficulty of the CST exams from two adjacent years. By subtracting the average student CST score from the previous grade from a predicted average CST score for the end of the current grade (derived from the mean and standard deviation of the previous cohort of students in that district), it becomes possible to estimate the average cross-year growth of the students taking the benchmark exam. Assuming that growth is linear through the year, it is then a simple matter to assign an average student CST score to the current benchmark by locating it on the growth trend-line that connects the previous year CST average with the predicted current year CST average (based on the performance of the previous student cohort). Multiplying the

standardized projected CST score for each student on the current benchmark exam by the estimated *non-standardized* mean and standard deviation for the current benchmark (based on where it falls on the cross-year trend-line), we can now assign an expected CST score, in the CST metric, to each student on each benchmark exam through the year. These expected CST scores permit researchers to track the individual growth of each student through the year, something which is not possible with raw benchmark percent correct scores.

As mentioned, this procedure relies on being able to calculate the relative difficulties of the CSTs from adjacent grades. Since California has opted so far not to equate its CSTs, we devised an *ad hoc* method for doing so. We studied the relationship between vertically equated scale scores nationwide and their mean grade equivalents (using data published by NWEA, creator of the RIT scale, and Metametrics, creator of the Lexile scale) to derive a likely growth curve of California students for reading and math. This curve shows high growth rates in the lower grades, steadily diminishing in the higher grades. We reinforced this with estimates of the percentage of students in California likely to show zero or negative growth between adjacent grades, anchoring them to a common curve. The position of zero-growth students in the cumulative normal distribution of standardized cross-grade differences then provides a way to estimate the relative difficulty of adjacent CSTs. The problem of assigning these cross-grade differences to a uniform metric is addressed by assuming that the size of the CST scale score unit is approximately uniform across grades, an assumption made more reasonable by the fact that the difference between Basic and Proficient has been defined to be equal for all grades, 50 scale score units.

All grade-level CST differences are then set relative to the Grade 6 definition of Proficient, having the effect of placing all the CSTs on a common vertical scale where a scale score of 350 corresponds to the Grade 6 definition of Proficient. We call this the “Grade 6 vertical scale” metric. To simplify the algorithm and handle situations where the student population has taken different CST exams in the previous year (common in upper grade math, for example), all CST scores are converted to the Grade 6 vertical scale prior to undergoing analysis. For reporting purposes, the Grade 6 vertical scale metric is converted to a “growth to expectation” metric defined such that every score is placed in relation to the state’s expectations of Proficient for the grades immediately above and below the student’s score on the Grade 6 vertical scale. To ease interpretability, the “Proficient” cut-point is defined to be at 75 plus the grade as a leading digit. Thus a value of 375 on the growth to expectation scale means Proficient on Grade 3 content; 775 means proficient on Grade 7 content. The growth to expectation scale ranges from 175 to 1275. The growth to expectation scale has a direct and intuitive appeal as a way to track students on a vertical scale that is adapted specifically to match the state’s definitions of Proficient for each grade. While its units are not equal interval but decline in size as a function of grade level, it does have a one-to-one monotonic correspondence with the ability-based equal interval Grade 6 vertical scale. This makes it suitable for research studies and growth measures, so long as students are compared only with other students of the same grade, and so long as it is remembered that the growth to expectation scale measures distance relative to state expectation rather than ability *per se*.

Also, like grade unit scales, growth to expectation scores are only relevant to the content that a student has already been taught. A 3rd grader who scores a 675 on the Grade 3 CST would probably not score a 675 on the Grade 6 CST. However, a 6th grader would probably score a 675 on the Grade 3 CST.

The EDS equating procedure was extended using a somewhat different methodology to include the General Math, Algebra I, Geometry, and Algebra II CST exams, as well as the CAHSEE exam, so that all are located on the same Grade 6 vertical scale. This allows the EDS growth to expectation scale to span Grades 2-11 for both Reading and Mathematics.

Measures on Individual Content Standards

So far we have only discussed how the benchmark exams are aligned and adjusted for difficulty in order to measure growth within the school year and across grades. NOUS also makes it possible to compute reasonably reliable measures at the level of individual content standards, even if they have as few as five items. It does this by using the entire benchmark exam data set to locate each student in a 2-dimensional space, then projecting that student's coordinate location onto the vector of each item, located in the same 2-dimensional space. This creates an expected score for each item that is much more precise and reliable than the student's raw score for that item. The same process takes place with the 1-dimensional Rasch model, but the expected student score on each item is essentially equivalent to the student's marginal logit score, so there is nothing to be gained by looking at expected scores for individual items. In the 2-dimensional case, however, each student's expected value is unique for each item, reflecting performance on the dimension of the vector embodied by that item. When these expected scores are averaged across the items in a content standard, we have a prediction of how each student is likely to perform on that content standard. As mentioned, this prediction, based on data drawn from the whole test, is much more precise than the average raw score of the items for that standard, roughly equivalent to the student having taken 20 items instead of five.

Measures on individual content standards are converted to a CST metric with its corresponding state-defined performance levels, making it possible to diagnose a student as Advanced, Proficient, Basic, Below Basic, or Far Below Basic on each content standard. All items and content standards are adjusted to have the same difficulty, which is defined in terms of the average score of all the students in the district. When converted to a CST metric, we would then say that the student's predicted CST score on a particular content standard is equivalent to what the student would have received were all the items in that standard of the same average difficulty as the test as a whole, an interpretation which might seem misleading for some standards that are unusually easy or difficult, but which is actually the *least* misleading way to diagnose genuine strengths and weaknesses on individual content standards at the student, classroom, and school levels, short of an official standard setting for each standard.

Applying a Benchmark Scaling Methodology

The table below shows the benchmark test results for a small sample of students in a southern California district. The test was administered to Grade 9 students in mathematics in spring 2007. Results in a “percent correct” metric are compared with results obtained using a scaling methodology for the test as a whole and for two content clusters on that test. The scaled results are presented both in a growth to expectation metric (here labeled “GTE”) and in an expected CST score metric. Performance level scores accompany each scale score, where 5 = Advanced, 4 = Proficient, 3 = Basic, 2 = Below Basic, and 1 = Far Below Basic. Standard error and reliability statistics are included, along with the number of items. “BM” stands for the Benchmark exam as a whole.

**Table 7: Grade 9 Mathematics Exam, Administered in Spring 2007 to All Students.
Scale scores reported relative to the Algebra I CST.**

Number of Items		80	12	8	.32 0.96	80	.	12	.	8	.
Standard Error		.	.	.		7	.	19	.	29	.
Reliability		.	.	.		0.99	.	0.91	.	0.78	.
Student	Grade	BM % Correct	Cluster1, % Correct	Cluster2, % Correct	GTE score	BM scale score	BM Perf. Level	Cluster1 Scale Score	Cluster1 Perf. Level	Cluster2 Scale Score	Cluster2 Perf. Level
A	9	0.54	0.58	0.63	846	339	3	339	3	334	3
B	9	0.50	0.58	0.25	813	318	3	318	3	323	3
C	9	0.49	0.58	0.63	772	314	3	316	3	294	2
D	9	0.81	0.92	0.88	1150	430	5	427	4	452	5
E	9	0.53	0.73	0.38	852	335	3	334	3	342	3
F	9	0.33	0.50	0.50	713	267	2	266	2	287	2
G	9	0.35	0.27	0.50	760	272	2	268	2	319	3
H	9	0.54	0.50	0.50	816	336	3	338	3	313	3
I	9	0.39	0.58	0.38	765	286	2	283	2	311	3
J	9	0.41	0.67	0.13	758	297	2	298	2	296	2
K	9	0.53	0.33	0.25	822	332	3	333	3	321	3
L	9	0.95	0.92	1.00	1224	473	5	470	5	497	5
M	9	0.40	0.67	0.38	737	291	2	291	2	285	2
N	9	0.71	0.75	0.13	1059	397	4	395	4	414	4
O	9	0.34	0.25	0.38	696	271	2	272	2	270	2
P	9	0.54	0.75	0.13	863	337	3	336	3	350	4
Q	9	0.71	0.75	0.88	1025	396	4	396	4	390	4
R	9	0.25	0.17	0.38	677	237	1	233	1	284	2
S	9	0.29	0.25	0.63	680	247	1	245	1	277	2
Mean		0.51	0.57	0.47	844	325	2.79	324	2.74	335	3.00

Table 7 highlights important differences between the benchmark raw score metric and a difficulty-adjusted scale score metric.

- **Comparison to State Expectations:** The raw percent correct metric sheds no light on how students are doing relative to state expectations. The scale score metric reports what each student would be expected to get on the Algebra I CST, with the corresponding performance level. We see, for instance, that a raw percent correct score of 0.51 corresponds in this case to an Algebra I CST scale score of 325, midway between Basic and Proficient. Expected CST results can be reported on any CST metric, as well as on the CAHSEE metric.
- **GTE scale:** The growth to expectation scale (GTE) reports where each student is relative to the proficiency levels defined for each grade. See that the average GTE score for this sample is 844, somewhat below the 875 that would correspond to “Proficient” on the Algebra I exam, well below the 975 that would correspond to “Proficient” on the Geometry exam (set by convention to define a Grade 9 proficiency target). The primary use of the GTE scale is to permit the measurement of growth across benchmark exams and grades on an interpretable vertical scale.
- **Correction for Cluster (or Standard) Difficulty:** The bold Mean statistics in the bottom row under “percent correct” would appear to indicate that students are performing more poorly on Cluster 2 (0.47) than Cluster 1 (0.57). The corresponding scale scores reveal that the situation is reversed. Students actually perform somewhat better on Cluster 2 (335) than Cluster 1 (324). The discrepancy is caused by several factors, the most important of which is that the raw cluster scores are not adjusted for item difficulty whereas the scale scores are. Thus, while a 0.47 looks low, when these students are compared with the rest of the students in the district (a proxy for cluster difficulty), they perform a little better than average on this cluster.
- **Correction for Aberrant Cluster Scores:** Because the raw percent correct metric does not take into account a student’s complete scoring pattern and consists of relatively few items, it can lead to results that misstate a student’s “true” ability on a particular cluster. Person N scores 13% correct on Cluster 2. His score on the benchmark as a whole is 71%, much higher. Do we trust Cluster 2 or his overall score more? In this case, the psychometric model assigned Person N a scale score of 414 on Cluster 2, above “Advanced” and above his score for Cluster 1. Person N’s total array of responses makes a 13% score on Cluster 2 highly unlikely.
- **Reliability Statistics:** The raw percent correct metric does not facilitate the calculation of standard error or reliability statistics. The GTE scale and expected CST scale does permit the calculation of such statistics, and tells us that the benchmark test is quite reliable on the whole (Reliability = 0.99, due to the large number of items) and that Cluster 2 is on the border of being reliable (Reliability = 0.78).

Figures 21 and 22 (following) illustrate the longitudinal nature of the growth to expectation scale.

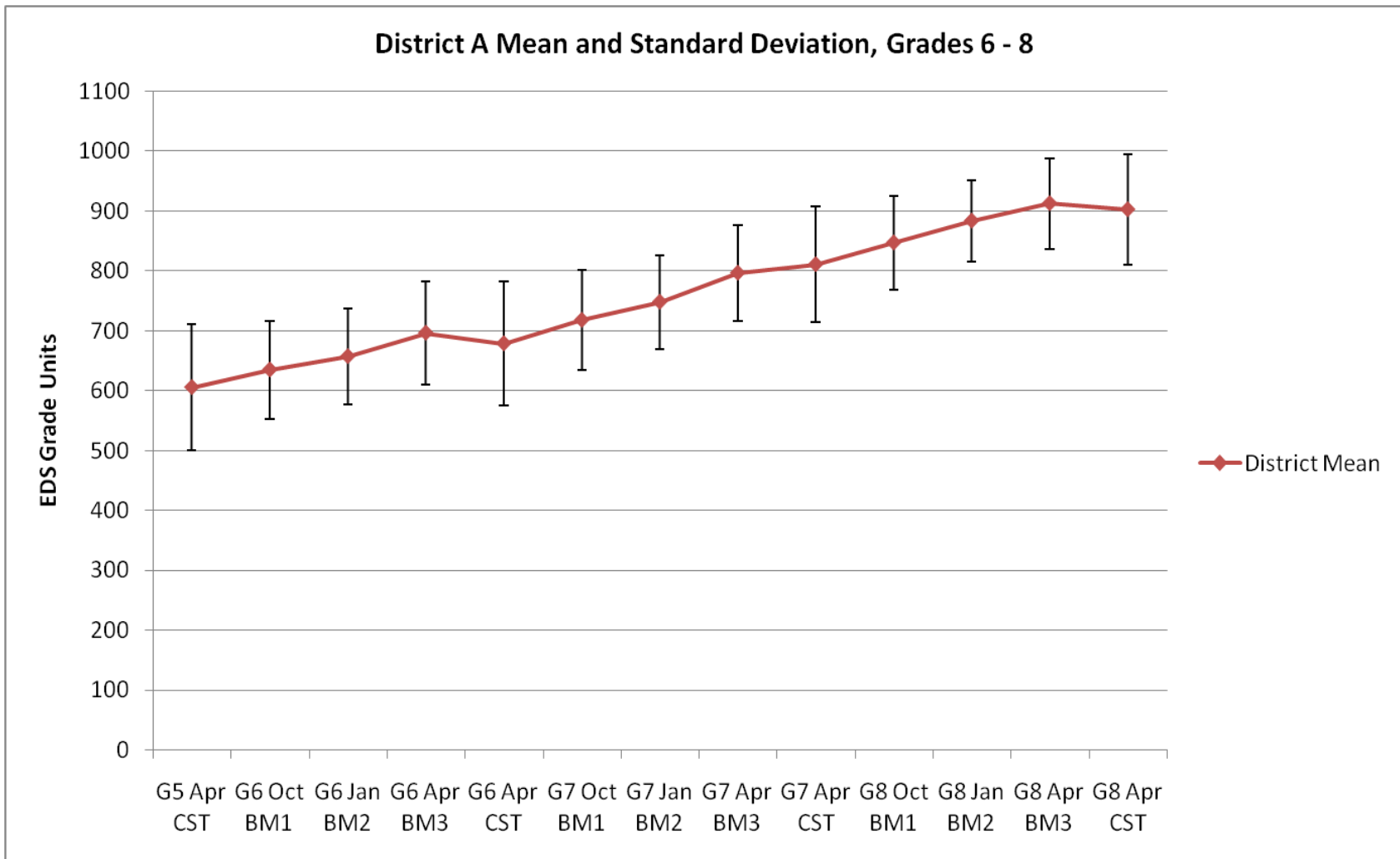


Figure 21: Average District Performance on Successive Benchmark and CST Exams, on the Growth to Expectation Scale

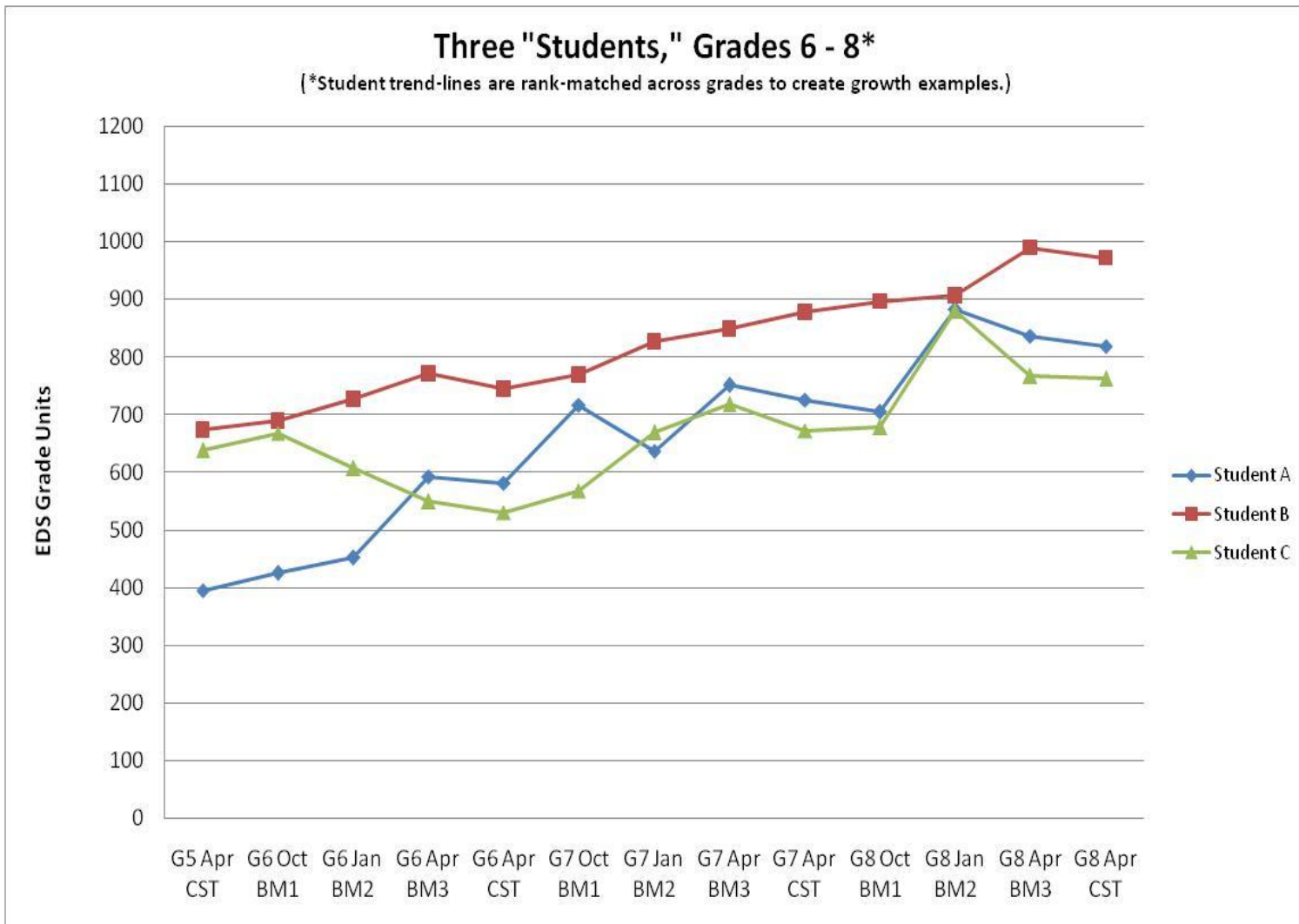


Figure 22: Individual Student Trend-lines

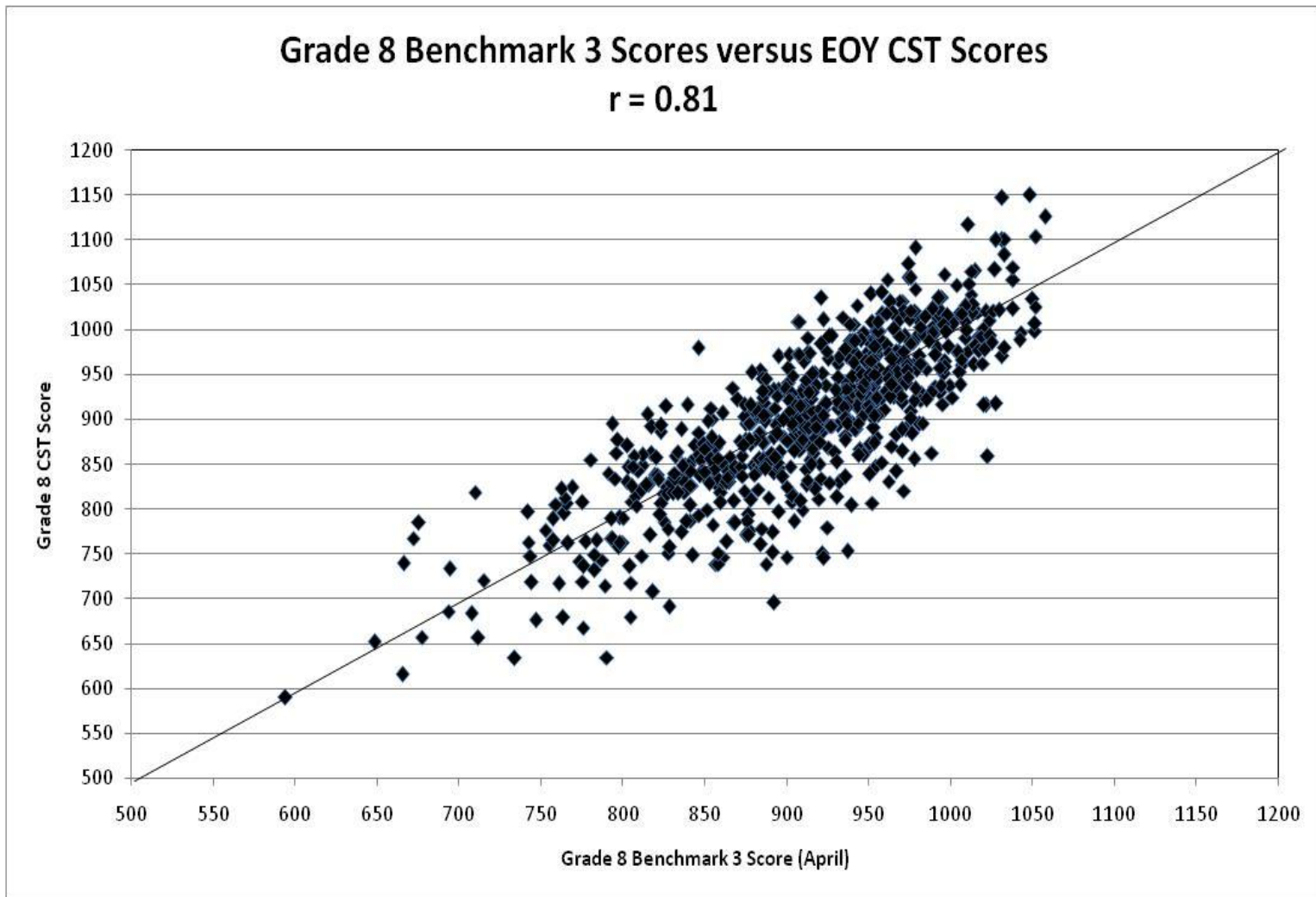


Figure 23: Relationship Between a Benchmark Exam and a CST Exam Administered in the Same Month, in GTE Units

Figure 21 illustrates the kind of longitudinal trend line that can be expected using the growth to expectation scale. Because the scale has been defined relative to State definitions of Proficiency, and because Proficiency levels tend to be set relative to expectations of what is reasonably possible of students in a grade level, these trend-lines tend to be fairly linear. On a vertical ability scale, one would expect to see trend-lines flattening out in the higher grades, a phenomenon well-known in the education field. By defining growth in terms of expectations rather than ability, we have not only oriented the scale relative to an official frame of reference, we have also straightened out a curved declining growth line to be in accord with what is reasonable in higher grades. An ability-based vertical scale is less arbitrary, but also less interpretable.

Figure 22 illustrates the same longitudinal trend-lines at the student level. (Due to the fact that we had only one year's worth of data for each student, we rank-matched students from adjoining grades to illustrate what a student-level longitudinal trend-line would look like.) The student trend-lines are, as one would expect, more erratic than district averages, but they are sufficiently coherent to reveal growth stories that could be informative to the teacher. For instance, a trend-line like that shown for Student B would indicate that Grade 6 was a rough year, but that the student's pace of learning picked up dramatically in Grade 7, though still subject to swings.

Validation

A true validation study of the EDS benchmark scaling method is outside the scope of this paper. Ideally, one would administer the CST exam along with each benchmark to observe the match between model predictions and actual scores and to plot trend-lines. Alternatively, one could create a simulated data set with a simulated CST exam to demonstrate the theoretical properties of the methodology.

While these kinds of validation studies are not currently feasible, Figure 23 does show the relationship between the Grade 8 CSTs (rescaled to a growth to expectation scale) and an EDS-scaled benchmark exam administered at approximately the same time. The 0.81 correlation between the two metrics, and the standard deviations of scores at each point in time, are about what one would expect given the measurement error of the CST and the measurement error of the benchmark exam. That, plus the proximity of the distribution to the identity line, suggests that the benchmark exam succeeds as a reasonable proxy of the CST exam when suitably equated.

Conclusion

In order for districts to meet accountability requirements, they need a way to evaluate the educational curricula, programs, and other educational factors hypothesized to affect gains. Program evaluation requires assessing the relationship between implementation of a program and the size of an achievement gain score over the same period. Without a way to calculate achievement gain scores, program evaluation is not valid. The CSTs, not being equated, are therefore not suitable for evaluating programs. However, when equated vertically and supplemented by similarly equated benchmark exams, gain scores can be calculated at the individual student level. Therefore, an equating procedure such as that used by EDS, along with corresponding data management support and instructional support, is important if schools and districts are to know which educational factors and programs are most effective and to act accordingly.

-----NOTES-----

PART IV: BRINGING PRACTICE TO POLICY

Over the past decade, statewide and national public school accountability systems have created a climate in which California schools and Local Education Agencies (LEAs) are under pressure to demonstrate high levels of success or at least meaningful gains towards higher levels of success. The convergence of California's Standardized Testing and Reporting Program (STAR) and Public Schools Accountability Act (PSAA) and the federal No Child Left Behind Act (NCLB) has created an environment of high stakes testing in the state. While the convergence of these programs and laws have no doubt had many positive impacts on the state's educational landscape, the fact that the spotlight of the state's testing environment is measuring academic achievement and growth at the school and subgroup level, rather than the individual student level remains an issue that needs to be addressed.

STAR Program and the CSTs

In 1997, the California legislation authorized the STAR Program through Senate Bill (SB) 376. SB 376 required students in grades 2-11 to be tested in English with State Board of Education (SBE) approved nationally Norm-Referenced Tests (NRTs) in reading, writing, and mathematics, with spelling added in grades 2-8 and history-social science added in grades 9-11. The same year, the SBE designated Stanford Achievement test, Ninth Edition, (Stanford 9) as the statewide pupil assessment.

The Stanford 9 was first administered in grades 2-11 in 1998. In 1998, the SBE authorized development of standards-based tests in English-language arts (ELA) and mathematics as augmentations to the Stanford 9. These standards-based tests were the genesis for all of the tests known as the California Standards Tests (CSTs).

In 2001, Senate Bill 233 reauthorized the STAR program for three additional years (2003-2005). Following the reauthorization of the STAR program, the SBE designated the California Achievement Tests, Sixth Edition Survey (CAT/6 Survey) to replace the Stanford 9. In 2003, all of the CSTs were separated from the Stanford 9 and included only questions written specifically for California's content standards. In 2004, Senate Bill 1448 extended the program through 2010, with the stipulation that the CAT/6 Survey be only administered in grades three and seven.

In 2005, Senate Bill 755 required that in addition to taking the designated STAR tests in English, Spanish-speaking English learners who either receive instruction in their primary language or have been enrolled in a school in the United States for less than 12 months are required to take a primary language test designated by the SBE.

The current STAR Program has four components: the CSTs, which are criterion-referenced tests that assess the California content standards in mathematics, English-language arts, science, and history-social science; the CAT/6 Survey, a nationally norm-referenced test; the California Alternate Performance Assessment (CAPA), an alternate

assessment to the CSTs that is designed to assess the performance of students with significant cognitive disabilities; and the Aprenda, La prueba de logros en español, Tercera edición (Aprenda 3), the designated primary language test in Spanish, a nationally norm-referenced test. Aprenda has since been replaced by the CDE developed Standardized Tests in Spanish (STS).

CST scale scores range from a low of 150 to a high of 600. There are five proficiency levels associated with the scale scores: Far Below Basic; Below Basic; Basic; Proficient; and Advanced. The scale for each subject test and grade level is centered on the Basic level, where the cut points are always 300 to 349. The cut points between Far Below Basic and Basic, as well as Proficient and Advanced, vary among grade levels and test subjects.

PSAA and the API

California's Public Schools Accountability Act (PSAA) of 1999 (Chapter 3, Statutes of 1999) authorized the creation of an accountability system for California schools with the two major focuses being school improvement and the measurement of academic achievement of all students. Provisions of the PSAA include the PSAA Advisory Committee, statewide evaluation, the Academic Performance Index (API), and the Alternative Accountability System for small schools and schools with non-traditional student populations, which is now under the Alternative Schools Accountability Model (ASAM). The three major components of the PSAA are the API, the Immediate Intervention/Underperforming Schools Program (II/USP), and the Governor's Performance Award (GPA) program.

The API is the foundation of the PSAA. Using a variety of measures of the testing results from the STAR Program and the California High School Exit Examination (CAHSEE), the API tracks the academic performance and growth of California's schools. While the PSAA law requires that test results constitute at least 60 percent of the API, currently test results constitute 100 percent of the API.

Based on statewide testing, the API is a numeric index given to schools and local education agencies (LEAs) that reflects performance level and is scored on a scale ranging from a low of 200 to a high of 1000. The statewide API performance target for all schools is currently 800.

The ongoing inclusion of new assessments necessitates that the API consists of two reporting cycles: Growth API and Base API. The Base API is the yardstick for comparisons with the Growth API. The 2007 Growth API results reported in August 2007 were based on students testing in spring 2007 and were calculated using the same methodology as 2006 Base API, which was reported in March 2007. The 2006 Base API was subtracted from the 2007 Growth API with the result being the 2006–07 API growth. Simply put, a school's current Base API is subtracted from the next year's Growth API to determine how much the school grew in a year.

In addition to reporting a Base API score, the Base API report includes a Statewide Rank (deciles 1 – 10), a Similar School Rank (deciles 1 – 10), an API Growth Target and an API target (Base API + Growth Target). Growth targets are set for each school and for each numerically significant subgroup in the school.

Numerically significant subgroups are defined as groups with 100 or more students with STAR Program test scores or groups with at least 50 STAR Program test scores that make up at least 15 percent of the school’s test scores. If they are numerically significant, the following subgroups can be included in API growth targets: African American or Black (not of Hispanic origin); American Indian or Alaska Native; Asian; Filipino; Hispanic or Latino; Pacific Islander; White (not of Hispanic origin); Socioeconomically Disadvantaged; English Learners; and Students with Disabilities. Schoolwide and subgroup Growth Targets depend on what their Base API scores were (see table below).

Table 8: Schoolwide Growth Target/Base API

	Schoolwide or Subgroup Base API			
	200 to 690	691 to 795	796 to 799	800 or more
Schoolwide or Subgroup Growth Target:	5% difference between Base API and 800	5-point gain	796 4-point gain 797 3-point gain 798 2-point gain 799 1-point gain	Maintain 800 or more

To meet state API targets, a school must equal or exceed its schoolwide growth target, and each numerically significant subgroup at the school must do the same. There can be up to 11 growth targets. At schools with 100 or more students enrolled in each content area prior to or on the California Basic Educational Data System (CBEDS) data collection date, at least 85 percent of the students need to participate in the testing. If that is not the case, then the API score is invalid.

Schools that meet the participation and growth criteria were originally eligible for monetary awards through the Governor’s Performance Award (GPA) Program but the program has not been funded since 2000-01. Now, through an extensive review process, they can apply to be classified as a California Distinguished School.

The PSSA mandates that schools that don’t meet growth targets or that are in the lower five API Statewide API Rank deciles are eligible for interventions through the Immediate Intervention/ Underperforming Schools Program (II/USP). The Quality Education Investment Act (QEIA) of 2006 assists schools ranked in either decile 1 or 2 as determined by the 2005 Base API.

No Child Left Behind and AYP

In January 2002, the NCLB Act of 2001 was passed by Congress. It changed the federal government's role in public education by requiring schools to demonstrate their success in terms of the academic achievement of every student. With students of greatest needs as the focus, NCLB emphasizes stronger accountability for results, expanded options for parents, and improving teacher quality.

As the largest federal program supporting elementary and secondary education, Title I of the NCLB Act is intended to help ensure that all children have the opportunity to obtain a high-quality education and to reach proficiency on state academic standards and assessments. Title I provides flexible funding that may be used to provide additional instructional staff, professional development, extended-time programs, and other strategies for raising student achievement in high-poverty schools.

NCLB includes four major requirements:

1. With academic content standards in place, states must test every student's progress toward those standards by using assessments that are aligned with the standards.
2. Each state, school, and LEA is expected to make Adequate Yearly Progress (AYP) toward meeting state standards of proficiency. Test results are sorted to measure the progress of all students; including numerically significant students who are economically disadvantaged, are from racial or ethnic subgroups, have disabilities, or have limited English proficiency. States commit to the goals of NCLB by participating in Title I. The primary goal of Title I is for all students to be proficient in English-language arts and mathematics, as determined by state assessments, by 2014.
3. State, school, and LEA performance is publicly reported in report cards.
4. If a Title I school or LEA fails to make AYP for two or more consecutive years in specific areas, it is identified for Program Improvement (PI). Schools or LEAs in PI must implement additional federal requirements.

Under NCLB criteria, schools and LEAs are required to meet or exceed criteria annually in four areas in order to make AYP: Participation Rate; Percent Proficient—Annual Measurable Objectives (AMOs); API as an Additional Indicator; and Graduation Rate (if applicable). There can be up to 46 targets that need to be met annually.

In order to comply with the AMO component, the California Department of Education (CDE) calculates the percent of students who scored proficient and advanced on the CST ELA and math tests at the school and LEA and subgroup level.

Limits of CST scores

The CST tests are the bedrock for measuring progress towards both the state API Growth targets and the federal NCLB AMOs. The CSTs accomplish the daunting tasks necessary to fulfill the measurement requirements of both accountability systems. However, they are less useful if one's desire is to track individual student growth. The STAR system and its CST tests were not designed for individual student assessment. Designed to assess schools, the CSTs say little about student performance for purposes of informing classroom practice and tracking student strengths and weaknesses.

A major weakness that has yet to be addressed is that the CSTs are not vertically calibrated. As was already noted, the cut points for the Far Below Basic, Below Basic, and Advanced levels differ by content area and grade. Because grades and content are scaled independently and different content standards are measured in different grades, one should not compare scale scores or proficiency levels across grade levels or content areas, though the practice is common. Not being vertically scaled, the CSTs cannot be used to measure individual cross-grade student growth, which seriously undercuts efforts to evaluate programs.

At the student level, the strand (aka cluster) scores within the subject area are the lowest level of analysis that one can attain. Usually there are five or six strands per subject. Strands are sometimes based on small numbers of items; therefore they may not be reliable or generalizable. The percentage correct of strands within the same test cannot be compared directly. Most notably, strands are not equated from year to year, so one can not compare the percent correct from year to year.

Because the CSTs are administered at the end of the school year, teachers and administrators are left in the dark about whether their students are on track to meet proficiency goals.

Policy Implications

The CST's are clearly limited when it comes to informing instruction and measuring individual student growth. However, it should be clear from this resource guide that powerful tools are available for proactive educational leaders to use local assessments to inform classroom instruction and predict outcomes on high stakes assessments. Benefits are readily attainable to districts and schools willing to commit to the following:

1. Build a sound local assessment based on standards reflected in the pacing guides of the local curriculum and quality item writing
2. Analyze the local assessment data using strong psychometrics
3. Validate local assessments through benchmark scaling
4. Implement instructional improvement
5. Predict high stakes assessment outcomes.

We invite you to take the next step forward with us towards creating a coherent assessment system that can provide meaningful feedback to improve student learning.

-----NOTES-----

FIGURES

1. Item performance	36
2. Three-phase model	37
3. Items sorted by difficulty (hardest to easiest)	39
4. Item 15	39
5. Tabular representation of function	40
6. Visual synectic and sentence frame	41
7. “Next Steps” worksheet	42
8. Students and items on same scale	43
9. Wright Map for Algebra I Math Winter 2008 Benchmark Exam	45
10. NRC Assessment Triangle	47
11. Possible Continua for Progress Maps	48
12. “Complexity of Functions” Progress Map	49
13. Wright Map for Complexity Construct Map	51
14. Springboard Item L1B	54
15. Springboard Item L1C	55
16. Springboard Item L1A	55
17. Springboard Item L5A	56
18. Performance levels based on Raw Scores	61
19. Performance levels based on Scaled Scores	62
20. Relationship Between Scaled Scores and Percent Correct	63
21. Average District Performance on Successive Exams On Growth to Expectation Scale	74
22. Individual Student Trend-lines	75
23. Relationship Between Benchmark Exam and a CST Exam Administered in the Same Month	76

TABLES

1. Advantages and Disadvantages of Item Types	17
2. Response Time Estimates by Item Type	18
3. Verb List	19
4. Item Allocation Planning Template	19
5. Item Calibration estimates, and fit statistics -- Springboard items	53
6. Percent correct across Level 1 Springboard items by grade level	59
7. Grade 9 Mathematics Exam Spring 2007, Scale Scores	72
8. Schoolwide Growth Target/Base API	81

REFERENCES

- Aiken, L.R. (2000). *Psychological testing and assessment (10th Edition)*. Boston, MA: Allyn Bacon
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.
- American Educational Research Association (AERA). (1999). *The Standards for Educational and Psychological Testing*. Washington D.C.: AERA
- American Federation of Teachers/National Council on Measurement in Education/National Education Association (AFT/NCME/NEA) (1990). *Standards for teacher competence in educational assessment of students*. Accessed online on 05/13/09 at www.unl.edu/buros/bimm/html/article3.html.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.
- Black, P. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education*, 5, 7-74.
- Bloom, B.S., et al. (1956). *Taxonomy of educational objectives. Handbook I: The Cognitive Domain*. New York: David McKay.
- Bloom, B.S., Madaus, G.F., & Hastings, J.T. (1981). *Evaluation to improve learning*. New York: McGraw-Hill.
- Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record*, 106, 429-458.
- Chhatterji, M. (2003). *Designing and using tools for educational assessment*. Boston, MA: Allyn and Bacon.
- College Board Springboard Program (2006). *College Board Standards for College Success*. Retrieved July 1, 2006 from <http://collegeboard.com/springboard>.
- Cronbach, L.J. (1990) *Essentials of Psychological Testing (5th Ed)*. New York: Harper & Row.
- Delgado, J. (2005). *Engaging strategies for all students: The Springboard example*. College Board Office of Research and Analysis. NY, New York.
- Erwin, T.D. (1991). *Assessing student learning and development*. San Francisco: Jossey-Bass.

- Ebel, R.L. & Frisbie, D.A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.
- Ericsson, K., & Simon, H. (May 1980). "Verbal reports as data". *Psychological Review* 87 (3): 215–251. [doi:10.1037/0033-295X.87.3.215](https://doi.org/10.1037/0033-295X.87.3.215).
- Ericsson, K., & Simon, H. (1987). "Verbal reports on thinking". in C. Faerch & G. Kasper (eds.). *Introspection in Second Language Research*. Clevedon, Avon: Multilingual Matters. pp. 24–54.
- Ericsson, K., & Simon, H. (1993). *Protocol Analysis: Verbal Reports as Data* (2nd ed. ed.). Boston: MIT Press.
- Farr, R., & Tone, B. (1994). *Portfolio and performance assessment: helping students evaluate their progress as readers and writers*. Forth Worth: Harcourt Brace.
- Gronlund, N.E. (2003). *Assessment of student achievement (7th Edition)*. Boston, MA: Allyn and Bacon.
- Haladyna, T.M. (1999). *Developing and validating multiple-choice test items*. Mahwah: Lawrence Erlbaum.
- Haladyna, T.M. & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 43-50.
- Illinois State Board of Education, Department of School Improvement Services. School and Student Assessment Section. (1995). *Effective Scoring Rubrics: A Guide to Their Development and Use*. Springfield, IL: Author.
- Johnson, D.W. & Johnson, R.T. (2002). *Meaningful assessment: A manageable and cooperative process*. Boston MA: Allyn and Bacon.
- Johnstone, C.J., Bottsford-Miller, N.A., & Thompson, S.J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved March 9, 2009, from the World Wide Web: <http://education.umn.edu/NCOE/OnlinePubs/Tech44/>
- Kennedy, C.A., Wilson, M., Draney, K., Tutuncuyan S., and Vorp, R. (2006). *ConstructMap software*. Berkeley Evaluation and Assessment Research (BEAR) Center. Berkeley, CA.
- Klein, S.P., Stecher, B.M. Shavelson, R.M., McCaffrey, D., Ormseth, T., Bell, R.M., Comfort, K., & Othman, A.R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, 11, 121-138.

- Lembo, John M. (1969). *The psychology of effective classroom instruction*. Columbus, OH: Charles E. Merrill Publishing Company.
- Lipton, & Wellman, 2004. *Data-driven dialogue: A facilitator's guide to collaborative inquiry*. Sherman, CT: MiraVia, LLC.
- McKeachie, W. J. (1999). *Teaching tips: Strategies, research, and theory for college and university teachers (10th Edition)*. Boston, MA: Houghton Mifflin Company.
- Meyer, J. & Land, R. (2003). *Threshold Concepts and Troublesome Knowledge: Linkages to Ways of Thinking and Practicing within the Discipline*. Report for Teaching and Learning Research Program. Edinburgh. Occasional Report 4.
- Moulton, M., & Mason, D. (2007). *State accountability vs. local testing: How benchmark exams can be turned into mini-CSTs*. California Education Research Association. Long Beach, CA.
- Murdock, J., Kamishke, E. & Kamischke, E. (2002). *Discovering algebra*. Emeryville, CA. Key Curriculum Press.
- National Research Council. (2001a). *Classroom assessment and the National Science Education Standards*. Committee on Classroom Assessment and the National Science Education Standards. J. M. Atkin, P. Black, & J. Coffey (Eds.). Center for Education, Division of Behavior and Social Sciences and Education. Washington, D.C.: National Academy Press.
- National Research Council. (2001b). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavior and Social Sciences and Education. Washington, D.C.: National Academy Press.
- Popham, W.J. (2000). *Modern educational measurement: Practical guidelines for educational leaders (3rd Edition)*. Boston, MA: Allyn and Bacon.
- Osterlind, S. J. (1998). *Constructing test items*. Boston: Klumer Academic.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. 4, 321-334.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago. University of Chicago Press (original work published 1960).
- Russell, Dale & Plakos, John. (1978). *Developing a pupil assessment system for proficiency-based instructional programs*. Office of the Los Angeles County Superintendent of Schools. Downey, CA.

- Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *The Handbook of Research on Teaching*, 4th Edition (pp. 1066-1101). Washington, DC: American Educational Research Association.
- Trice, A.D. (2000). *A handbook of classroom assessment*. New York: Addison Wesley Longman, Inc.
- Vygotsky, L. S. (1978). *Mind in society: the development of higher psychological processes*. Cambridge: Harvard University Press.
- William, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education*, 11, 49-65.
- Wilmot, D. B. (2008). *Assessing progress toward college readiness with psychometric and cognitive models of student learning in mathematics*. Doctoral dissertation (University of California, Berkeley).
- Wilson, M. & Scalise, K. (2003). Reporting progress to parents and others: Beyond grades. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom*. NSTA Press: Arlington, VA, 89-108.
- Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13(2), 181-208.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Wright, Benjamin D., & Masters, Geoffrey N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA PRESS.