Validating a Learning Progression in Mathematical Functions for College Readiness

Diana Bernbaum Wilmot, Alan Schoenfeld, Mark Wilson, Danielle Champney & William Zahner

University of California, Berkeley
University of California, Santa Cruz

Available online: 06 Oct 2011

To cite this article: Diana Bernbaum Wilmot, Alan Schoenfeld, Mark Wilson, Danielle Champney & William Zahner (2011): Validating a Learning Progression in Mathematical Functions for College Readiness, Mathematical Thinking and Learning, 13:4, 259-291

To link to this article: http://dx.doi.org/10.1080/10986065.2011.608344

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Validating a Learning Progression in Mathematical Functions for College Readiness

Diana Bernbaum Wilmot, Alan Schoenfeld, Mark Wilson, and Danielle Champney

University of California, Berkeley

William Zahner

University of California, Santa Cruz

Current college admissions tests inadequately predict college success (Geiser & Studley, 2002) and provide insufficient information for students and teachers regarding college readiness at earlier stages of students’ academic careers (Olson, 2006). This article reports on the creation and validation of a more informative test rooted in college readiness standards in mathematics. Items and scoring guides were created using mixed-methods of assessment design, and assessment results were analyzed and validated using item response theory, student interviews, and teacher evaluations of test items. This research shows how an innovative measurement approach can be used for more accurately assessing college readiness, and reveals one possible means of providing better measures along a central strand of mathematical knowledge required for students entering college.

In the United States, mathematics remediation rates at the university level have prompted educators and policymakers to propose new college readiness standards and have highlighted the need for alignment between college entrance exams and K-12 mathematics curricula. In June 2010, governors and state education officials from 49 US states and territories, together with Achieve, Inc., the College Board, and the ACT®, a nonprofit organization known for producing a commonly used college entrance exam across the United States, released the Common Core Standards for Mathematics and English-Language Arts. The Standards highlight “connections among topics that are most important for success regardless of a students’ pathway” through the educational system (Common Core Standards Initiative, 2009). Thus, it is our job as educators.

Correspondence should be sent to Diana Bernbaum Wilmot, University of California, Berkeley, 347B W Rincon Ave, Campbell, CA, 95008, USA. E-mail: dianajoy@alum.berkeley.edu
to provide our students with a range of experiences that enable them to make connections. This includes the development of innovative assessments to measure such patterns and connections. This study, while not connected directly with the new Standards, illustrates how to develop and validate predictive assessments based in domain-specific theories of student learning such as those that guide the writing of standards, more broadly.

Recent work on learning progressions in K-12 education (Corcoran, Mosher, & Rogat, 2009; Clements & Sarama, 2009; Sarama & Clements, 2009) promise to advance our understanding of how to model growth and capture student trajectories so that teachers can use assessment data to inform curricular and instructional decisions (Kennedy & Wilson, 2007). In this article, we have started to define one learning progression that can measure a trajectory of student’s ability to make connections across representations of mathematical functions from sixth through twelfth grade. While previous research has investigated how students develop an understanding of mathematical functions (Kaput, 1992, Leinhardt, Zaslavsky, & Stein, 1990), a learning progression in this content area has yet to be validated. In addition, we have designed an assessment to measure this development and gathered evidence to capture the range of student progress from 2356 students in California, Connecticut, New Jersey, and Massachusetts.

Following the four principles of the Berkeley Evaluation and Assessment Research (BEAR) Assessment System (Wilson & Scalise, 2003), the set of assessments for this research was designed to measure a developmental path of learning, align with college readiness standards, integrate with current practices in curriculum development and instruction, and compile evidence of validity and reliability using both qualitative and quantitative measures. This kind of research requires a collaborative effort across disciplines, including input from cognitive scientists, measurement experts, policymakers, mathematicians, and mathematics educators. In addition to producing high-quality assessments, this work demonstrates the potential for further collaboration among researchers in these fields.

While a number of high-quality items currently exist on various mathematics and/or college readiness assessments, this research suggests a new method for structuring and scoring students’ work on formative mathematics assessments. The validity and reliability associated with scoring guides developed in this research are documented through the consistency of multiple raters, alignment of student think-aloud interviews and written responses, a standard psychometric model, and interviews with teachers at several levels of the educational system. The items and scoring guides were developed to validate a cognitive framework, resulting in assessments and scoring guides that help students make a smooth transition from pre-secondary mathematics to college-level mathematics.

BACKGROUND

Major standard-setting organizations such as the College Board (2006), the National Council of Teachers of Mathematics (2000, 2006), and the University of California (2003) and California State University Academic Senate (1997) all suggest that entering college students should be able to make connections across multiple representations of mathematical functions. A literature review produced by Berkeley Futures Project summarized and elaborated several important aspects of understanding mathematical functions that are necessary for college success. The review also explicated a developmental theory of student understanding in this domain based on prior research of students’ understanding of mathematical functions. Three primary findings from
that review of the literature were: (1) Working with functions as algebraic rules relies on students’
development of prerequisite skills such as identifying and working with variables (Leinhardt
et al., 1990; Kieran, 1992); (2) Initial student understandings of functional relationships rely on
single representations and local interpretations (e.g., “function as rule,” recursive table reading,
or point-wise graph interpretation reading) (Kaput, 1992; Friel, Curcio, & Bright, 2001); and
(3) Students who competently solve problems with functions connect multiple representations of
functions, focus on local and global features of a graph, and flexibly move between the process
(or rule-based) and object representations of functions (Moschkovich, Schoenfeld, & Arcavi,
1993; Chazan & Yerushalmy, 2003; DeMarois & Tall, 1999).

Current statewide standardized tests for accountability and college admissions tests, such
as the SAT®, ACT®, and AP® examinations, include items to assess students’ knowledge of
mathematical functions. While these items sometimes address aspects of the important findings
about functions listed previously, they are most often scored quantitatively, according to the
amount of correctness demonstrated by a student’s response to the item. For example, a scoring
rubric (Figure 2) for a standard item from the New Jersey Mathematics Curriculum Framework
(Figure 1) does not connect students’ responses to a model of the students’ mathematical under-
standing of the functional relationship. As is typical, the focus of the sample rubric is on the
amount of correctness rather than the level or type of understanding demonstrated by a particu-
lar response. Because no developmental theory underlies the scoring rubric, students’ numerical
scores will not expose the differing levels of mathematical understanding necessary to produce
such scores. While an item may appear to tap into students’ higher-level thinking (e.g., asking
the student to show his or her work, explain, predict, etc.), the description underlying the scor-
ing of the students’ work is ultimately what determines the construct being measured (Wilson
& Sloane, 2000). Greeno, Pearson, and Schoenfeld (1996) have argued that assessments that mea-
sure students’ proficiency from a cognitive perspective are important for moving mathematics
education research forward. They further argued that the measurement community has not kept
up with either the changes in mathematics curriculum or the cognitive research in mathematics
learning that has been ongoing for more than three decades.

In this research we created an assessment composed of multi-level items (such as the Hexagon
Pattern item in Appendix A; student work shown will be referenced later) and used appropriate
measurement techniques such as item bundling to measure the higher-level connections that are
not typically measured by current college admissions assessments. Student responses to items
on the developed assessment were assigned codes to reflect the cognitive levels in the proposed
learning progression rather than context-less numerical scores. Thus, the levels used for scor-
ing were designed to highlight students’ development of understanding rather than amount of
correctness. This new approach is designed to be more formative than traditional assessments,
because the quantitative results provide meaningful feedback about student thinking. Therefore,
this article focuses on the following research questions:

1. Can a learning progression based on a review of the literature in the area of mathemat-
ical functions be designed to measure students’ progress in making connections across
representations of mathematical functions?
2. How can researchers establish reliability and construct validity to verify the learning
progression used in the college readiness assessment?
3. How do teachers plan to use the assessment results to make curricular and instructional
decisions?

1.) For each of the figures below, the length of one side of a square is 1 unit.

1 square

2 squares

3 squares

a.) Find the perimeter of one square, two squares connected along an edge, and three squares connected along their edges. Make a table of values to show how the perimeter changes with respect to number of squares.

<table>
<thead>
<tr>
<th>Number of Squares</th>
<th>Perimeter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b.) Determine a function rule to describe the pattern above. Show all work.

c.) Predict the perimeter for 10 squares. Show all work.

FIGURE 1 A problem from the New Jersey curriculum framework (copyright permission).

RESEARCH DESIGN

With a four-phase data collection and analysis process, including both qualitative and quantitative measures, this study followed a sequential, developmental mixed-methods approach (Greene & Carcelli, 1997). Each data collection phase was followed by analysis and writing to inform the next phase of data collection and analysis. The four phases were:

I. Literature review, expert paneling, and development of learning progression
II. Item development, cognitive (a priori) think-aloud interviews, and item piloting
III. Assessment design, administration of the assessment, post-hoc think-aloud interviews, score moderation sessions, and final scoring of items
IV. Interviews with middle school and high school mathematics teachers about their interpretations of the assessment report.
This mixed methods approach follows Wilson’s (2005) recommendations for establishing evidence for validity and reliability in developing an assessment. Following an iterative process, the data were analyzed at each phase in the data collection cycle to triangulate the findings (Greene & Carcelli, 1997). The methods and results are described in separate sections, corresponding to the four phases of the overall research design.
A panel of mathematics, policy, measurement, and science experts, funded by Berkeley Futures Project, met monthly during the 2006–2007 academic year to review the literature related to college readiness standards and policies concerning college readiness. The group of experts decided to focus the content of the assessment around a significant topic—making connections across multiple representations of mathematical functions—that is applicable to numerous college majors (e.g., architecture design, computer programming, business, demographics, social policy, economics, women’s studies, and mathematics). The panel determined that “functions” and functional relationships are a core algebraic concept that spans kindergarten through twelfth grade (National Council of Teachers of Mathematics, 2000) and constitutes a significant focus of collegiate mathematics. This finding was based on recommendations from groups such as the California State University Academic Senate (1997), which recommended that entering college students understand “various representations of functions—using graphs, tables, variables, words,” because students have to interpret mathematical functions across multiple disciplines. The group also focused on functions because there is a solid knowledge base of student learning in this domain. Researchers have spent decades studying how students think about mathematical functions (Piaget et al., 1977; Leinhardt et al., 1990; Sfard, 1992; Zachariades, Christou, & Papageorgiou, 2002; Moschkovich et al., 1993; DeMarois & Tall, 1999). This research was used as a basis for the development of the learning progression.

Developing the Learning Progression

A construct can broadly be considered as a “theoretical object of interest in the respondent” (Wilson, 2005, p. 13). A “learning progression” or “construct map” refers to a unidimensional latent variable by which we can display a continuum that describes the varying levels of a particular construct. The critical features of a learning progression or construct map are thoughtful, substantive definitions of both the phenomenon being measured and of the proposed levels of the construct as well as the idea of an underlying continuum that is evident in both ordering of the respondents and the ordering of item responses.

The functions learning progression designed for this research was built from the SOLO taxonomy (Biggs & Collis, 1982), a generalized assessment framework that can be used to measure students’ ability to make connections within and across domains (Biggs, 1999). We argue that the SOLO taxonomy can be applied to the area of mathematical functions, with few amendments, since our goal was to measure students’ ability to make connections across multiple representations of mathematical functions.

The general SOLO taxonomy includes five levels of student understanding. We adapted the SOLO taxonomy in the functions learning progression by adding a sixth level called “Prealgebraic.” The version of the functions learning progression used in this research is in Figure 3. It describes student responses as a progression, read from bottom to top, representing increasing sophistication of student understanding of mathematical functions.

---

1Complete treatment of the development of the Functions Learning Progression is found in Bernbaum & Champney (2008). This paper includes only a brief synopsis of the development.
### Functions Learning Progression:  
Making Connections Across Multiple Representations of Mathematical Functions

<table>
<thead>
<tr>
<th>Levels of Sophistication:</th>
<th>Student is able to:</th>
<th>Response to items</th>
</tr>
</thead>
</table>
| **Extended Abstract (5)** | Make connections not only within the given subject area, but also beyond it, able to generalize and transfer the principles and ideas underlying the specific instance. | - Predict, explain, and synthesize their understanding to a real-world context.  
- Solve non-routine problems including non-algorithmic functions. |
| **Relational (4)** | Demonstrate understanding of the significance of the parts in relation to the whole. | - Compare/contrast information given in multiple representations of functions to demonstrate understanding of the content.  
- Recognize which representations to choose for the context.  
- Select representations and move fluently among them to achieve a solution. |
| **Multistructural (3)** | Make a number of connections, but the meta-connections between them are missed, as is their significance for the whole. | - Make connections across more than two representations (e.g. symbolic, tabular, graphic and verbal representations).  
- Recognize more than one relevant feature of a functional relationship. |
| **Unistructural (2)** | Make simple and obvious connections, but their significance is not grasped | - Make pair-wise connections between representations. |
| **Prestructural (1)** | Acquire bits of unconnected information, which have no organization and make no sense. | - Interpret graphs where both variables have to be interpreted, or where time as independent variable.  
- Demonstrate understanding of function in one representation. |
| **Prealgebraic (0)** | Acquire prerequisite skills | Demonstrate understanding of:  
- functional dependence, where a change in one variable affects another variable.  
- continuous variables (time, distance) and dichotomous variables (hot or cold).  
- Relative order and measurement to define variables. |

**FIGURE 3** Functions learning progression.

### Levels of the Learning Progression

At the lowest level, the Prealgebraic level, students demonstrate that they are learning prerequisite skills to work with a function to solve problems (Leinhardt et al., 1990). Students at this level do not use variables to express functional relationships between quantities, and they appear to be learning this prerequisite skill.

When students answer items at the second level (Prestructural), they address one representation of a function at a time and make no connection with other representations. For example, in a tabular function, where \( x \) and \( y \) are represented by values in a vertical table, Kaput (1992) suggested that students may initially only understand the recursive numerical pattern, “reading
down the table.” Such a response indicates that the students may not understand how $x$ and $y$ are related numerically, and creating an explicit function (where the changes in $y$ can be explained as a function of the changes in $x$) is likely out of reach for these students.

Students performing at the next level, Unistructural, are beginning to make connections between representations, but those connections are relatively simple. For example, students may be able to translate from a tabular function to a symbolic function but unable to describe the function verbally (DeMarois & Tall, 1999).

Students at the Multistructural level are able to make a number of connections between representations of functions. At this level students can connect the tabular, symbolic, and verbal representations of a function and recognize more than one relevant feature of a functional relationship. For example a response at the Multistructural level would demonstrate that the student can explain how the slope-parameter in the symbolic form of a linear function relates to the common difference between subsequent $y$-values in a well-ordered table (Zachariades et al., 2002).

At the Relational level, students compare and contrast information given in multiple representations of functions to demonstrate understanding of the content. In addition, they may be able to decide which representations are most appropriate for a given context, use several representations, and move fluently among them to achieve a solution (Moschkovich et al., 1993; Chazan & Yerushalmy, 2003).

Finally, at the most sophisticated Extended Abstract level, students use and apply algebraic functions to solve problems given in a real-world context and predict, explain, and synthesize their understanding by solving or creating novel problems. At this level, students work with functions in a nonalgorithmic way (i.e., going beyond applying memorized rules), which is considered difficult for most students (Sfard, 1992). It is important to note that while the functions learning progression is a framework for interpreting responses, it does not suggest that students will learn mathematical functions in a lock-step fashion sequentially following these proposed levels.

METHODS AND RESULTS: PHASE II

Item Development

The CRA included several different item types: traditional multiple-choice items from the College Board Springboard (2006) program and three types of open-ended items (all of which were developed during the Berkeley Futures Project): multiple choice with justification, constructed response, and meaning equivalence. Multiple choice with justification items asked students to choose one of the four answer choices and then explain why they chose that answer. Constructed response items required students to write their responses with words, algebraic formulas, and/or numbers. Meaning equivalence items were similar to multiple-choice items in that there were four answer choices, but for some items more than one answer was correct and students were instructed to choose all the correct answers.2

2The meaning equivalence item type was chosen because it was designed to measure if students can recognize equivalence among multiple representations across various content areas (Shafrir, 1999).
Twelve constructed response multi-level items were developed for this assessment. In this section we describe the process of item development by focusing on one item (Hexagon Pattern item in Appendix A), because the development process for other items on the CRA was similar. To capture the range of student thinking on this item we used a think-aloud protocol and asked a diverse group of students to work on the problem. Doing think-aloud problem-solving interviews prior to full-scale data collection allowed us to refine the items and the functions learning progression before administering a final version of the written CRA in Phase III of the study.

Participants

To determine how test takers reason through the items and calibrate the scoring guides, the think-aloud interviews were conducted in March 2004 with four “expert” students who had all completed two years of secondary algebra. The process was repeated in June–July 2007, with 11 students that ranged from US Grade 6 to their fourth year of college. The sample included students taking courses such as pre-algebra, algebra II, trigonometry, pre-calculus, advanced placement calculus BC, discrete mathematics, and mathematics for elementary school teachers.

Think-Aloud Protocol

After the think-aloud protocol was read aloud, students were given two problems unrelated to the proposed assessment to practice thinking aloud. During the interview, the researcher wrote field notes documenting particular strategies and behaviors that could not be captured on the audiotape and reminded the participant to talk about what he or she was thinking whenever he became silent. The same protocol was used for all participants to ensure comparability in responses. The set of 11 interview transcripts was considered as collective evidence of students’ understanding of mathematical functions, and thus was a resource for identifying levels of understanding across a wide range of abilities.

Coding and Analysis of Think-Aloud Interviews

Analysis of the think-aloud interview transcripts proceeded in two phases, beginning with descriptive coding and followed by thematic coding. Because the functions learning progression was used to guide the development of the scoring guide, researchers used the functions learning progression for the thematic analysis. We first coded all items individually by looking for levels of proficiency across students in order to help establish evidence toward validity of internal structure. We then looked at the thematic codes across items to determine if there was reasonable evidence to suggest the functions learning progression as a viable tool for interpreting student responses across multiple items.

Analysis entailed iterative cycles of learning progression development and identification of both longitudinal and comparative patterns. The first author reviewed collections of item responses and drafted case memos to describe the range of thinking by students across each
item. Matrices were also used to compare patterns of individual student responses across the set of items. The first author discussed and challenged preliminary interpretations of the item responses and used descriptive coding to capture material relevant to interpretation of student work (e.g., learning progression refinement, scoring, analysis of whole student and whole item) as well as factors impacting or interfering with learning across time. Reports of the coded material were number coded by grade level to facilitate analysis of patterns of change.

The analysis of these data helps evaluate the soundness of the learning progressions and the items, but most importantly, the findings from the think-aloud interviews help inform the revision of the scoring guide, which can be used to develop a reporting system for teachers (Wilson & Sloane, 2000).

To demonstrate how we designed an item that maps onto multiple levels of a learning progression, we turn to the Hexagon Pattern item (Appendix A—Note how this item is similar to the squares item from the New Jersey Framework in Figure 1 and to a released NAEP item; http://nces.ed.gov/nationsreportcard/itmrls/). This item was adapted from Balanced Assessment/MARS Middle Grades (2000), but the scoring rubric was completely revamped for this study. The Hexagon Pattern item asks students to interpret a geometric pattern that represents the perimeter of a chain of hexagons. There are five parts to this item:

1. Complete a tabular representation of the pattern of the perimeter.
2. Describe and explain a verbal representation of that pattern.
3. Create a symbolic representation of that pattern.
4. Describe and explain a verbal representation of a pattern of the perimeter for a chain of $n$-gons.
5. Create a symbolic representation of that pattern, for $n$-gons.

Parts 1, 2, and 3 are included to evaluate students’ ability to reach the prestructural, unistructural, or multistructural level on the functions learning progression. Parts 4 and 5 extend the previous parts of the item, giving students the opportunity to relate their knowledge about hexagons to $n$-gons, thereby extending their knowledge to a non-routine problem and reaching the relational and extended abstract levels of the functions learning progression.

To demonstrate how the development of the Hexagon Pattern item influenced the quality and interpretability of students’ cognitive processes, five students’ think-aloud responses to Part 2 are described next (students EV, DS, LT, KS, and NK). After each student’s process for solving the Hexagon Pattern item is described, each student is characterized, and a summary of the student’s responses highlights the diversity of responses at the end of this section.

EV, a sixth grader who noticed “connected sides” during Part 1, explained that “some people would think that [the perimeter for 100 hexagons] would be 600, but had you gone up, each one in the middle is only up 4, then each one at the end you’re adding 5.” He proceeded to talk through the multiplication problem of 98 times 4 and said confidently, “So I think [the answer] would be 402.” The basic approach highlights that each of the end hexagons account for five sides each, while all other hexagons contribute four units to the perimeter.

3The first iteration of this item occurred during item piloting on a 2004 WestEd project called the Learning Zone.
DS, a tenth grader in Algebra II looked at the picture and tried to find a numerical pattern. DS wrote “5, 4, 5” and then “5, 4, 4, 5” and then “5, 4, 4, 4, 5” representing the perimeter of 3, 4, and 5 hexagons in a chain. When recording a formula for Part 3, he recalled that he should “plus the number of hexagons . . . um . . . multiplied by 4 I guess.” Despite encouragement to describe the pattern verbally, DS still wanted to use $10 + 4n$ as the expression that described the pattern.

LT, an eleventh grader, looked at the picture to determine the growth of the geometric pattern. She said,

So if it was um, 10 hexagons, you would have 60 total sides and subtract two for each, because have 2 that you can’t see of 10 hexagons, so that would be 20 sides that you can’t see. But then you’d have to subtract 2 from that but only for the 1 on each end of the chain. So subtract 2 sides, you get 18 sides. So I’d subtract that from 60 and that would give me hopefully my answer of 42.

KS, a twelfth grader, used a “common difference” rule that he determined in Part 1 and wrote, “You would begin with your given hexagon and add four each time. You would take the number of hexagons then multiply by 4 and add 2 to get your perimeter.”

Finally, NK, a college sophomore, wrote out his description in words, “5 (2 end hexagons) + 4 (# of middle hexagons).”

As evidenced by the diversity of student responses, there were a variety of ways to address this problem. The think-aloud interview with DS on this problem led into a scaffolded conversation to help him solve the problem while EV, KS, LT, and NK were all able to describe the pattern in their own words. In each case described, none of the students described the process simply as “adding four,” thus researchers were able to detect if students’ verbal descriptions from Part 2 were connected to the algebraic formulas they provided in Part 3. These think-aloud interviews suggest that correct responses to Parts 2 and 3 will indicate that students have reached the multistructural level on the functions learning progression.

To summarize the complete think-aloud sessions with the students, LT and NK reached the extended abstract level on the functions learning progression because they were able to generalize their formula to a non-routine problem without a given algorithm. DS reached the relational level because, while he was able to describe the process for finding the perimeter of a chain of $n$-gons verbally, he was unable to generate a formula. KS was able to reach the multistructural level by showing his ability to make connections between the tabular, verbal, and symbolic functions, but it was clear from his struggle to find a formula for $n$-gons that he was not able to extend those connections. Finally, EV reached the unistructural level on functions learning progression because he was only able to make a connection between the tabular and verbal representations of functions. Recall that the purpose of the think-aloud interviews was to refine the items and the functions learning progression so that they could be used and validated with a larger assessment sample. The descriptions provide a basis for some of that refinement, since the student work and reasoning could be tied to the functions learning progression on a smaller scale to inform how items might best be amended before moving the CRA to large-scale distribution.

Next we describe Phase III of this research, administering the revised CRA widely and validating it using multiple data sources including a psychometric model, post-hoc think-aloud interviews with selected test-takers, and teacher interviews.
METHODS AND RESULTS: PHASE III

Assessment Design

Because of time restrictions and the desire to not overwhelm young students with difficult items, a series of assessments (referred to as Forms A, B, C, D, E, F, and G) was designed with a set of commonly linked items. The linking procedure with test equating calibration and anchoring, described in more detail in Wilmot (2008), was established to make valid comparisons of student proficiency across six grade levels.

Participants

After refining the items in the CRA based on the a priori think-aloud interviews, the assessment was administered to 2356 students in 125 classrooms in US Grade 6 through Grade 12. Table 1 demonstrates the range of student demographics across the sample. Assessment administration occurred November–December 2007. The teachers in each class were provided with the same set of directions for administration, which was read verbatim to the class. Students were given 40 minutes of class time to complete the assessment.

Post-Hoc Think-Aloud Interviews

A select group of students\(^4\) participated in post-hoc think-aloud interviews conducted after students completed the written version of the CRA. The students were taken to a quiet room in their school and were asked to recall what they were thinking as they solved the open-ended items. The interviews lasted 10–15 minutes and generally included only one of the open-ended items on the CRA. In total, 21 post-hoc interviews were conducted across all 12 open-ended items. These interviews were summarized and analyzed by two researchers to check for agreement between students’ written responses on the open-ended items and their verbalized thought process.

Coding Students’ Written Responses

To score the finished assessments, three researchers participated in training sessions focused on creating robust scoring guides that used the levels on the functions learning progression, across items. Since the aim of scoring guides is to have a usable framework for interpreting student test results, robustness requires that a score for a response on one item (i.e., EA, for Extended Abstract) could intuitively have a similar meaning in terms of students’ demonstrated level of performance when it is used as a score on a different item.

Responses to each of the open-ended items were assigned a code corresponding to the various levels on the functions learning progression. Due to the large sample and extensive time required

\(^4\)The first author asked teachers to choose students who could verbalize their thoughts, but were also representative of the range of proficiency levels in their class. Students were also chosen based on the items they answered on the test.
### TABLE 1
Range of Demographics Across Sample

<table>
<thead>
<tr>
<th>Type of School</th>
<th>Charter-Low Income</th>
<th>Comprehensive Public</th>
<th>Private Elite Boarding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location of School</td>
<td>California</td>
<td>California</td>
<td>New Jersey</td>
</tr>
<tr>
<td>Founded</td>
<td>2005</td>
<td>1954</td>
<td>1994</td>
</tr>
<tr>
<td>Grade Levels</td>
<td>6–9</td>
<td>9–12</td>
<td>9–12</td>
</tr>
<tr>
<td>Size of School</td>
<td>195</td>
<td>1687</td>
<td>3080</td>
</tr>
<tr>
<td>% Free Reduced Lunch/ Financial Aid</td>
<td>59%²</td>
<td>17%</td>
<td>14%</td>
</tr>
<tr>
<td>Demographics of School</td>
<td>African American:</td>
<td></td>
<td>African American:</td>
</tr>
<tr>
<td></td>
<td>61%</td>
<td>4%</td>
<td>7%</td>
</tr>
<tr>
<td></td>
<td>Hispanic:</td>
<td></td>
<td>Hispanic:</td>
</tr>
<tr>
<td></td>
<td>29%</td>
<td>23%</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>Asian:</td>
<td></td>
<td>Asian:</td>
</tr>
<tr>
<td></td>
<td>4%</td>
<td>16%,</td>
<td>17%</td>
</tr>
<tr>
<td></td>
<td>Caucasian:</td>
<td></td>
<td>Caucasian:</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>50%</td>
<td>65%</td>
</tr>
<tr>
<td></td>
<td>Other:</td>
<td></td>
<td>Other:</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>% ESL (English not first language)</td>
<td>38%³</td>
<td>11%</td>
<td>1%</td>
</tr>
<tr>
<td>Number of students in Sample</td>
<td>182</td>
<td>207</td>
<td>508</td>
</tr>
</tbody>
</table>

² Based on 2006–2007 year, which only included 6–8 graders.
³ Based on 2006–2007 year, which only included 6–8 graders.

To score students’ written responses, 20% of the papers from each classroom were randomly selected for scoring. To evaluate inter-rater reliability, the open-ended items were scored with a 10% overlap of scoring across the three researchers. These scores were quantified and then analyzed using the Rasch Model (Rasch, 1961, 1980) to establish reliability and validity (see Empirical Evidence).

### Score Moderation Sessions

Five score moderation sessions took place between January and March 2008 to train two additional researchers in the scoring process (described in more detail in Wilmot, 2008). Each moderation session ended with a separate final scoring guide for each of the items, all coordinated with the functions learning progression (Figure 3). The scoring guide for the Hexagon Pattern item is in Appendix B and all of the scoring guides are described in Wilmot (2008). The accuracy and the consistency of the scores across raters and across the functions learning progression will be addressed next.
Empirical Evidence

This section documents the empirical evidence to support the cognitive theories hypothesized in the learning progression as well as the design of the items and the scoring guides. To gather evidence that the CRA can describe students’ progress accurately and consistently, the results provided in this section aim to evaluate the learning progression, items, and scoring guides with high psychometric standards of validity and reliability.

For the purpose of this research, reliability evidence includes inter-rater reliability and internal consistency indicators such as Cronbach’s Alpha (Cronbach, 1990) and person separation reliability (Wright & Masters, 1982). Validity is investigated by looking for evidence based on response processes and internal structure (Wilson, 2005).

Reliability Evidence

Inter-rater reliability is the degree to which raters agree in the scoring of the items. A level of agreement was calculated across all items by calculating the difference between the scores from Raters 1 and 2; a level of agreement of “0,” indicates complete agreement in the scores, and a level of agreement of “1” indicates that Rater 1 scored the item response one level higher than Rater 2. As indicated in Figure 4, the level of agreement ranges from −3 to 3, and the difference between the scores of Rater 1 and Rater 2 was “0” across 207 out of 277 items scored by two raters. This represents an exact agreement in scores for approximately 75% of the data. The remaining 70 items scored show an agreement within at least two levels from the rubric. This level of agreement is evidence that the scoring guides were being utilized in a consistent way and suggests that the functions learning progression offers a reasonable framework that can be consistently applied across items designed to measure students’ ability to make connections across multiple representations of mathematical functions.

The number of items double scored along with the correlation between the rater scores is presented in Table 2. Eleven of the twelve items show statistically significant correlations between the scores of the two raters, with 10 of the items indicating correlations above 0.75, demonstrating a reasonable level of consistency in scoring across graders for those items. Only the scores...
TABLE 2

Correlation Between Scores for Rater 1 and Rater 2 on Open-Ended Items

<table>
<thead>
<tr>
<th>Name of Item</th>
<th>Number of Items Double Scored</th>
<th>Correlation Between Scores of Rater 1 and Rater 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hexagon Pattern</td>
<td>69</td>
<td>0.93**</td>
</tr>
<tr>
<td>Crude Oil—100 years</td>
<td>33</td>
<td>0.69**</td>
</tr>
<tr>
<td>Educational Accomplishments</td>
<td>22</td>
<td>0.72**</td>
</tr>
<tr>
<td>Gender Gap</td>
<td>19</td>
<td>0.50*</td>
</tr>
<tr>
<td>Parallel Functions</td>
<td>17</td>
<td>1.00**</td>
</tr>
<tr>
<td>Intersection Function</td>
<td>26</td>
<td>0.81**</td>
</tr>
<tr>
<td>Equivalent Functions</td>
<td>35</td>
<td>0.99**</td>
</tr>
<tr>
<td>Functions with same x-intercept</td>
<td>16</td>
<td>0.84**</td>
</tr>
<tr>
<td>Functions cross x-axis</td>
<td>9</td>
<td>0.37</td>
</tr>
<tr>
<td>Staircase Toothpick</td>
<td>20</td>
<td>0.94**</td>
</tr>
<tr>
<td>Postage Stamp</td>
<td>8</td>
<td>0.77*</td>
</tr>
<tr>
<td>Crude Oil—24 Hours</td>
<td>9</td>
<td>0.76**</td>
</tr>
</tbody>
</table>

**statistically significant at the .01 level.
*statistically significant at the .05 level.

associated with two items on the assessment have a surprisingly low correlation between Rater 1 and Rater 2. Possible reasons for this low correlation include lack of clarity in the scoring guides and more attention needed during training and moderation. This finding is discussed in the full report of the research (Wilmot, 2008).

Cronbach’s alpha, calculated from the correlations between pairs of items, is one indication of consistency among items (Cronbach, 1990). In this case, Cronbach’s alpha is a measure of the consistency of the scoring guide. It follows that the consistency of the scoring guide allows us to make a conclusion about the consistency of the functions learning progression. There are no absolute standards for what is acceptable, but by convention, an Alpha equal to or greater than 0.60 is considered a minimum acceptable level in low-stakes tests, while some authorities argue for a stronger standard of at least 0.70.5 The Cronbach’s alpha of 0.77 across open-ended items on the CRA suggests an adequate statistic for internal consistency.

Person separation reliability, calculated as a function of student ability and item difficulty, describes how well the items on an assessment can differentiate between students of various ability levels (Wilson, 2005). A statistic closest to 1 indicates that a large proportion of variance in student ability that can be accounted for by the items included on the test. On this test, the person separation reliability was 0.70 for the open-ended items, suggesting that the open-ended items do a reasonable job distinguishing between students of various proficiency levels described on the functions learning progression.

5For high stakes tests, the reliability coefficients on the 2008 California Standards Test ranged from 0.85-0.94 (CST Technical Manual, 2008).
Validity Evidence Based on Response Processes

Some cognitive scientists argue that it is difficult to capture students’ cognitive thought processes when they are expected to write their thoughts on paper (Nisbett & Wilson, 1977). In the post-hoc think-aloud interviews, students were asked to describe their thought process while solving problems. Interview transcripts were used to score the student responses during these interviews. The new score based on the interviews allowed researchers to compare the students’ verbal description of their thought process with their previously written response to the same item. The results of response processes, described next, provide validity evidence of the CRA.

A level of agreement between written and verbal responses, shown in Table 3, was calculated across all 21 interviews by examining the difference between the scores on the written and verbal responses by the same student. A difference of zero between the scores on the written and verbal responses indicates complete agreement, suggesting that the student showed the same level of performance on the written assessment and during the think-aloud interview. If the written response was scored at least one level above the verbal response (e.g., extended abstract level (5) compared to the relational level (4)), this would indicate that the written response showed a higher level of performance than the verbal response. The reverse is also true: when the difference is negative, the verbal response was better able to show a high level of performance than the written response. A negative difference is what one would expect to find most often (Nisbett & Wilson, 1977).

Approximately 62% of the post-hoc think-aloud interviews showed zero difference between the two scores (Figure 5), illustrating that these written items, together with the scoring guides and functions learning progression, capture approximately the same level of student performance as a verbal interview.

The level of agreement across all the items (Figure 5) ranges from −2 to 2, suggesting that the written and verbal responses differed by no more than two levels on the functions learning progression. The average difference is −0.2 across all the items tested, suggesting that there may be a tendency toward the negative (i.e., verbal response may indicate higher levels of performance than written response). The level of agreement was negative for 5 of the 21 interviews. While

<table>
<thead>
<tr>
<th>Score on Written Response—Score on Verbal Response</th>
<th>Level of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Complete agreement</td>
</tr>
<tr>
<td>1</td>
<td>Captured higher cognitive thought process on written response</td>
</tr>
<tr>
<td>−1</td>
<td>Captured higher cognitive thought process on verbal response</td>
</tr>
</tbody>
</table>

TABLE 3
Level of Agreement Between Written Response and Post-Hoc Think-Aloud Interviews

Downloaded by [University of California, Berkeley], [Mark Wilson] at 12:52 11 October 2011
this sample size may be small, the evidence may further support Nisbett and Wilson’s (1977) findings.

Validity Evidence Based on Internal Structure

The functions learning progression, based largely on the SOLO Taxonomy (Biggs & Collis, 1982), has been hypothesized to measure students’ college readiness as six developmental levels in the area of mathematical functions (Figure 3). Creation of this intentional structure in the learning progression (i.e., attempting to make a score of unistructural have the same meaning across items) allows researchers to use a measurement model to determine if the empirical results agree with the theory hypothesized (Wilson, 2005). The following results suggest the cognitive framework is developmental, but the difference between levels across items is not uniform enough to formalize cut points.

While multiple measurement theories abound for application to linked and calibrated assessment data (e.g., Generalizability theory, 2-parameter or 3-parameter Item Response Modeling, etc.), we chose to use Rasch-based modeling (Rasch, 1961, 1980) because the results provide a meaningful interpretation to help teachers use the assessment data in a formative way (Wilson, 2005) (described in more detail in Phase IV). The Rasch model (1961) provides a convenient way to develop estimates of student proficiency and item difficulty using the same scale based on probability of observed responses. This measurement model allows us to analyze the developmental nature of the learning progression through a visual interpretive map known as the Wright Map (Wright & Masters, 1982). As described in more detail in the teacher interviews, the Wright Map, in conjunction with a learning progression, provides a strong criterion-referenced interpretation of student proficiency (Wilson, 2005). The mathematical equation used in this measurement model is described in Appendix C.
Comparing Theory to Measurement Model

As one might expect, there was some variation by problem in the number of students classified at different levels. This is, most likely, a function of differences in task difficulty and student background (e.g., students found linear functions the easiest to deal with), and the classifications reflected this. Of course, such variation is typical of testing; that is why the process of equating scores (which only judges the amount of learning) was created.

To check the consistency and distinction of this learning progression, we examine the Wright Map in Figure 6. This map shows a visual interpretation of the estimated student proficiency (on the left side) and the estimated item difficulty (on the right side) after calibrating the items. The interpretation of the Wright Map is meaningful psychometrically because each of the thresholds on the Wright Map relate to the levels described on the functions learning progression. We can use the Wright Map in Figure 6 as a summative tool to validate our framework and as a formative tool to refine the instrument itself.

The evidence for the theoretical framework proposed in the learning progression is based on the contrast between the levels proposed in the functions learning progression and the empirical

---

6The Xs on the left hand side of the map represent the proficiency of 688 students as distributed across the sample. Since there is not enough room on the Wright map to show all 12 items, 8 of the 12 items are represented on the right side, with their respective titles written across the top of the map. The Equivalent Functions item is similar to the missing four items on the map.
data illustrated in the Wright Map. Based on the way we built the functions learning progression (Figure 3), we expect that the levels of the construct will arise in a certain order (PA on the bottom, then PS, US, MS, R, and EA on the top). The estimated order for each of these levels in the Wright Map suggests a concordance between the theoretical expectations in the learning progression and the empirical results in the Wright Map. Thus, the theoretical framework as a developmental learning progression holds true.

However, the estimated locations of each of the levels on the Wright Map are not consistent across items. For example, the calibrated item parameters on the Wright Map (Figure 6) suggest that it is easier for students to reach the EA level on the Hexagon Pattern item than on any other item. The same students who have a 50% chance of answering the Hexagon Pattern item at the EA level have a 12% chance of reaching the EA level on the Educational Accompishments item, and a negligible chance of reaching the EA level on the Crude Oil items. This is probably due to the way in which the Hexagon Pattern item was vetted and written to ensure students’ opportunity to reach the R and EA levels. A sample of student work on the Hexagon Pattern item (Appendix A) illustrates one of the unique ways students can apply their knowledge about hexagons to a polygon with any number of sides. Such examples triangulate the findings from the original think-aloud interviews used for item development (Bernbaum & Champney, 2008) and demonstrate how students’ understanding of multiple representations of mathematical functions in the problem (i.e., tables, equations, and words) can help them generalize their knowledge in ways that may not be as accessible in other items.

In the Hexagon Pattern item, students can reach the multistructural level by completing the table, finding a rule for the pattern of values in the table, and explaining the process for finding the perimeter of 100 hexagons. Students can use a guess and check method on this item by plugging in values until they find an equation that works. In contrast, students cannot work backwards or substitute values to reach the MS level on the Educational Accompishments item, where they must be able to recognize the graph as two linear functions on the same axes and determine a rate of change for any time period on the graph. It was even more challenging for students to reach the MS level on the Staircase Toothpick and Gender Gap items because they require knowledge of a quadratic function.

Cut Points

Examination of the Wright Map (Figure 6) implies that there is insufficient evidence to make cut points that illustrate the six levels of the functions learning progression. The levels in the Hexagon Pattern item suggest collapsing levels of the functions learning progression into three categories (i.e., PS/PA, US/MS, R/EA), but this trend is not apparent in all the items. In contrast to the Hexagon Pattern item, the Educational Accompishments item (available upon request from first author) does the best job differentiating between students at all levels described on the functions learning progression. The levels shown in the Wright Map (Figure 6) are sufficiently spaced and clearly distinct.

Despite the lack of differentiation for all six levels, we can see from the empirical results using the functions learning progression that students who are able to make connections between multiple representations of mathematical functions (multistructural) are at a higher proficiency level than students who are just acquiring the prerequisite skills to interpret one representation at a time.
(unistructural). When working with linear functions, students make connections between a symbolic representation and a tabular representation more easily than they make connections between a graphical representation and verbal representation. Students had most difficulty making connections between a graphical representation, a tabular representation, and a verbal representation of the same function. Thus, it appears that students who are able to detect a linear relationship from a table of values may also be able to recognize a linear relationship as it is presented in a graph, equation, and words. However, the reverse may not be true—students who can identify a linear relationship in an equation may not be able to recognize linearity in a graph or in a table of values. These findings suggest that according to current college readiness standards, students who are preparing for college ought to, at the very least, recognize linear relationships in three different types of representations: symbols, graphs, and tables.

Next Steps

Using the Wright Map as a formative tool to refine the instrument, our next step will be to examine the scoring rubrics and samples of student work where the differentiation between student responses is unclear (e.g., see levels PA, PS, and US on the Equivalent Functions item in Figure 6). Since this instrument is the first of its kind, and no previous data have been calibrated, we should not be surprised to find that the empirical data in the Wright Map (Figure 6) and the levels on the functions learning progression are not perfectly aligned. While the levels are not uniform across all items, there are at least three levels of distinction in every item. This is a reasonable starting point for the next iteration of the CRA’s development.

Using Wilson’s (2005) suggestions for instrument refinement, our next steps include (a) examining the scoring guides that have been developed for each of the items, (b) investigating the match between the items developed and the functions learning progression, and (c) revisiting the original theory proposed in the functions learning progression. The results presented here provide useful validity evidence to inform future iterations of the instrument, aligned with each of these next steps.

METHODS AND RESULTS: PHASE IV

Participants

In summer 2008, two middle school teachers (SS and ST) and two high school teachers (BY and JS) who were involved in Phase III of the research were interviewed in pairs about the assessment results. The purpose of these focus group interviews was to share the results of the assessment, clarify the significance of the information reported, and discuss how teachers might use the information in the results to make curricular and/or instructional decisions. The interviews lasted about 45 minutes each and were video recorded.

Interview Protocol

An interview protocol was designed to structure the conversation around the ways in which the teachers might use the information from the report of the CRA results to make curricular
and instructional decisions. Follow-up questions at the end of the interview focused on how to improve the assessment reporting for future dissemination.

Report of the CRA Results

A report of the CRA results (R-CRA) indicating student performance on the CRA was designed to offer a strong criterion-referenced interpretation of student proficiency and allow the teachers a chance to see samples of student work at various levels. Each school/district participating in the study received an R-CRA in June 2008. Depending on the number of different mathematics courses that were included in the analysis, the R-CRA ranged from 15–27 pages. Providing such a detailed report to the participants at this point in the development allowed us to identify which types of reporting structures (i.e., bulleted lists, tables, samples of student work, Wright Maps, etc.) could help teachers interpret the data and engage in a meaningful dialogue around how they may use the information to make curricular and/or instructional decisions.

Data Analysis

Teacher interviews were analyzed with descriptive and thematic coding. Summary statements were compared across two researchers to ensure triangulation of the findings.

1. Descriptive codes were used to locate topically similar talk. For example, the relevant talk included responses to particular interview questions, such as decisions about possible changes in curriculum and/or assessment. All the interviews were coded by looking for potential decisions that may come as a result of the information from the assessment.

2. Because vertical articulation of the curriculum is expected to be an outcome of the assessment results, grade levels or math course levels were used for thematic analysis. These thematic codes were used across the interviews to determine if there was reasonable evidence to suggest that the assessment results may be a viable tool for understanding the development of students’ understanding of mathematical functions across US grades six through twelve.

3. Summary statements were drafted regarding grade level, and course level curriculum and matrices were used to summarize teacher responses within and across interviews. The analysis of this data helps evaluate the use of the assessment and, most importantly, the findings from the interviews help both inform the revision of a usable reporting system for teachers and provide recommendations regarding college admissions criteria in mathematics.

The results from the teacher interviews regarding instructional and curricular decisions are discussed next. The teacher interviews are also used as a lens for analyzing the relationship between student understanding of mathematical functions and college readiness.

Teacher Interview Results

Teachers who were interviewed explained that the results in the R-CRA “tell us where [our students] are” and “what our curriculum should do to meet them” (Wilmot, Champney, Zahner,
Acknowledging the formative nature of this assessment, teachers in one school saw the R-CRA (excerpt in Appendix D) as a way to:

- **Check student progress**: “This provides a lot of data to see that we are doing a good job moving those [kids] you know sixth to eighth [grade].” (ST)
- **Make instructional decisions**: “Now I have the data to back up my gut that I wanted to spend time reviewing [with ninth graders] just to get them up to speed.” (SS)
- **See learning as a trajectory**: “This really helped me feel like okay, here’s the trajectory . . . we want to make sure that [a student’s learning] trajectory keeps going up, rather than flattens out.” (SS)
- **Plan curriculum across years**: “We need to make sure we do the same thing from eighth grade to twelfth grade . . . think about planning curriculum for a whole group, instead of just five kids.” (ST)

In addition, the middle school math teachers indicated that they could interpret the results to understand the content that their students are actively learning. For example, when comparing the performance of their eighth graders to their ninth graders using the results (Appendix D), one teacher (SS) said, “that’s pretty much exactly what I expected to see.” The report showed that the eighth graders, on average, scored better than the students in other classes at her school. SS went on to remark that the eighth graders had outperformed the sixth, seventh, and ninth graders on every test they have taken thus far (including the statewide tests), offering further evidence to the validity of these data. In explaining the low performance of the ninth graders, SS noted that the ninth graders were a smaller group of students that had not experienced stability at the school like the students in other grades had. As the first class of students to enter a new charter school three years prior, the student enrollment within that cohort had exhibited more turnover than the other classes.

The remainder of this section documents teachers’ consideration of curricular and instructional decisions in light of the assessment results. The decisions presented next are merely considerations; at this juncture, it is unknown if the districts and schools have implemented any changes discussed during these interviews. The decisions are divided into four levels of the school system: classroom, department, school, and district.

### Classroom-Level Decisions

After looking at samples of student work on the Hexagon Pattern item and average student performance in algebra II on the Wright Map (excerpt in Appendix E), one teacher (BY) suggested the use of “out-of-the-box kind of problems,” to help build students’ capacity to generalize from the concrete to the abstract. As teachers at their school are in the process of adopting a new textbook, another teacher (JS) said that their decisions would be made in light of the R-CRA.

### Math Department-Level Decisions

After looking at samples of student work on the Hexagon Pattern item and the average student performance on the Wright Map (Appendix D), SS and ST considered the introduction of \( n \)-gons into the curriculum the following year, so they could “push their students further” and
keep their students’ learning “trajectory on a steady [incline] going up.” While they considered the teaching of abstract concepts as “pushing them further,” the teachers used the Wright Map (Appendix D) as evidence that the eighth graders were “ready to move to the next level.” Using the R-CRA as their guide for discussion, they expressed the need to consider curriculum that goes beyond concrete examples while also building a foundation in the sixth and seventh grade mathematics curriculum, which could be accomplished by focusing on the connections between representations—they said the curriculum should focus on the “overlap of information . . . that you can get from one representation, but you can’t get from another representation.”

However, when considering the development of the middle school mathematics curriculum, ST said he wouldn’t change the way higher order concepts are introduced “after the curriculum is complete.” He said, “[we need to be] working on them developmentally . . . putting in all the foundational stuff so that they can—either abstractly kind of put their mind to it and see it visually . . . I think it takes the whole, pretty much a good portion of the year to get them ready in that spot.”

**School-Level Curricular Decisions**

JS and BY indicated that at their comprehensive high school there was a schoolwide focus on expository reading across the content areas. In light of the assessment results presented with samples of student work from the Hexagon Pattern item for honors trigonometry students, JS suggested that it might be necessary to also focus on expository writing across the disciplines. JS commented, “You can be the smartest person in the world, you can have the best knowledge base of anyone, but if you can’t communicate, it’s worthless.” JS thus used the R-CRA to argue that communication should be taught across all subjects.

**District-Level Curricular Decisions**

Both BY and JS felt that there has been a decline in the skills of algebra II students in the recent years. The assessment results (Appendix E) confirmed these thoughts for BY. She recognized that average double block algebra I students (students at-risk of failing algebra I) were actively learning the same material as her algebra II students. That is, both groups of students were still learning the prerequisite knowledge in order to make connections across representations of mathematical functions. While this may be expected for the double block algebra I students, BY expected her algebra II students to have mastered the prerequisite skills. Having taught algebra II for nearly 10 years, BY blamed the decline in the skills of her algebra II students on the institutional pressure to teach Algebra in the eighth grade, and the lack of emphasis on algebraic concepts during the geometry curriculum. JS agreed, saying “[students from eighth grade algebra] have a very, very weak basis when they get to algebra II, they have a weak basis in geometry, and we’re just setting them up for failure.” As a solution, they suggested a more integrated approach to the study of geometry and a more rigorous study of algebra in the eighth grade.

Looking across the themes, the teachers reported that the R-CRA proved useful as a formative measure to help them check student progress, view learning as a trajectory, plan curriculum...
across years, and make instructional and curricular decisions at all levels of the system. However, the strength of this finding is tempered by the limited sample of teachers interviewed. While an R-CRA was generated for all seven schools in the original study (Wilmot, 2008), formal follow-up interviews were only conducted with four teachers at two schools. Further, the four teachers included in this study were willing participants and understood the purpose of the assessment as a formative tool. These teachers may also not be typical in their ability to interpret statistical data because they were credentialed to teach secondary mathematics.

Champney (2010) further explored teachers’ interpretations of graphics such as those in appendices D and E, with a much wider sample of approximately 50 teachers. That study concluded that an average teacher without one-on-one instruction from a member of the research team has a reasonable chance of interpreting the Wright Map by comparing, most often, the distribution of students or the middle portion (by grade) with the content map on the far right hand side. Average teachers had a harder time coordinating across all three columns of the figures or extracting the information to make instructional or curricular decisions.

Teachers across disciplines (e.g., English language arts, mathematics, science, social science) and grade levels (K-12) have begun engaging with Wright Maps and formative assessment results in a very meaningful way (Duckor et al., 2010). The results from this study have helped inform the development of professional training for teachers and administrators to use formative assessment results as a lens to understand student thinking, which thereby helps teachers use data to inform instruction and curricular decisions.

CONCLUSIONS AND IMPLICATIONS

There is a very real need for high-quality formative assessments that provide meaningful feedback to students. We have developed a formative measure that captures what students know and provides the tools for moving students forward and placing them appropriately. Knowing how much they know (the current state of the art) does not help to do that, as the massive failure rate and need for remediation indicates. The measure we have built moves the field solidly toward productive formative assessment.

This research explored the creation of a cognitive framework based on the SOLO taxonomy and a review of the literature in the area of mathematical functions. The empirical results of the open-ended items on the CRA suggest that such a cognitive framework can be used to measure students’ progress in making connections across representations of mathematical functions, but the construct may contain three distinct levels rather than six. It is difficult to know if the results suggest faults within the cognitive framework, limitations of the item design, or discrepancies across the scoring guides. Like most instruments, the College Readiness Assessment must be developed through an iterative process to refine the cognitive framework, items, and scoring guides (Wilson, 2005). The developmental theory behind the College Readiness Assessment, detailed in the functions learning progression, proved to be a robust framework for creating scoring guides, which could consistently be utilized across multiple items.

The investigation of reliability suggests that the scoring guides are reasonably clear for raters to score assessments in a consistent manner. The validity based on response processes suggests that the items are capturing students’ proficiency in making connections across multiple representations of functions in a written assessment. The validity based on internal structure suggests
that there is a hierarchy, or a learning progression, in describing the way students make connections across multiple representations of functions. However, some items should be further vetted by experts and revised again to give students the opportunity to reach the R and EA levels on the functions learning progression, as they may offer more extensions than originally proposed.

The research reported here only focuses on one specific area—students making connections in mathematical functions. While functions are an important topic that spans much of the content in school mathematics, there are other important areas in mathematics that may be meaningfully linked to college readiness (California State University Academic Senate, 1997; Common Core Standards Initiative, 2009; National Council of Teachers of Mathematics, 2000). This narrow focus on connections between representations of mathematical functions is an unavoidable limitation of this study. However, the findings presented here open the door for larger scale work in the design of instruments to measure college readiness across multiple areas of school mathematics.

The intensive training sessions involved in scoring 12 open-ended items required more time and manpower than initially anticipated. Additionally, evaluating these types of assessments required expert judgment from mathematically competent scorers (all scorers had college degrees in mathematics). This indicates that test administrators must be willing to pay the price for the meaningful information provided by a robust assessment like the instrument described in this article. While this research engaged graduate student researchers, future research should include opportunities for middle and high school mathematics teachers to participate in scoring the assessment in order to build their capacity to develop, interpret, and use formative assessments to drive instruction (similar to an effort by the Silicon Valley Math Initiative, funded by the Noyce Foundation [Foster & Noyce, 2004]). As a result, the next iteration of this research (in progress) will document and measure a learning progression for how teachers build assessment expertise from pre-service training to in-service professional development. Teachers’ assessment expertise will be developed in a series of professional development modules that emphasize using assessment results as a lens to understand how students think, and then inform subsequent instruction.

We would expect that as teachers learn more about the content of the CRA, student test scores would improve because teachers would likely alter their curriculum and instruction to reflect the content included in the CRA (Schoenfeld, personal communication, December 4, 2008). One might surmise that when the assessment actually contains meaningful content, then “teaching to the test” is not a bad idea. However, the CRA developed here may be an existence proof rather than a set of instruments (i.e., multiple versions of the same test) that could be reliably used in practice. To have a collection of instruments that could be used as a consistent and constant measure would require an extensive item bank and calibration of equivalence between those items.

Although the area of focus for this research—mathematical functions as it pertains to college readiness—has been deemed an important strand of learning by many sources, this study has not specifically tied “college readiness” as measured by the assessment with college success measured by grades, graduation rates, or some other metric. Future research will include a secondary data analysis of placement exams to investigate the relationship between students’ performance on the CRA and their grades in college. Furthermore, interviews with math professors from Phase IV (not reported here) offer further recommendations for revised college entrance requirements in future publications (Wilmot, Champney, Zahner, Wilson, & Schoenfeld, in preparation).
This research provides a demonstration of the sorts of interpretation, framework, and items that may aid the development of admissions tests at universities in the United States, aimed at evaluating the content set forth in the new standards. For universities interested in revising mathematics admissions requirements by trying to provide criteria that are more descriptive, rather than a simple list of courses, this research provides an example by describing students’ college readiness as it develops from middle school through high school in one important area of mathematics.

This research provides solid evidence documenting a high-quality formative measure of what students know, enabling teachers to help move students forward and to place them appropriately. Absent such a “forward-looking test” that provides meaningful feedback to students that aligns with the expectations of college-level preparedness, massive failure rates, and the need for remediation at the college level are likely to continue. Students must be a part of the solution, but the mechanism for helping them do well is to provide early indicators, diagnostic information, and worthwhile feedback. This research provides a step in that direction.

ACKNOWLEDGEMENTS

This research was supported by a Berkeley Futures grant through the University of California Chancellor’s Office. The Berkeley Futures grant involves a collaborative effort among researchers across disciplines and departments at the University of California, Berkeley. The findings and opinions represented in this report do not necessarily represent the views of the University of California Chancellor’s Office. We gratefully acknowledge the contributions of Mark Wilson, David Stern, Angelica Stacey, Robert Jorgensen, Emiliano Gomez, and Karen Chang.

REFERENCES


7The University of California Office of the President (UCOP) has established a UC Task Force to revisit the recommendations of the Academic Senate regarding the description of mathematics admissions requirements.


APPENDIX A: HEXAGON PATTERN ITEM\(^1\) (SAMPLE OF “EA” STUDENT WORK)

1. For the following geometric pattern, there is a chain of regular hexagons (meaning all 6 sides are equal):

\[
6n - 2(n-1) \quad \text{perimeter} = 6n - 2(n-1)
\]

\[
6n - 2(n-1) = 1\text{ hexagon}
\]
\[
6n - 2(n-1) = 2\text{ hexagons}
\]
\[
6n - 2(n-1) = 3\text{ hexagons}
\]
\[
6n - 2(n-1) = 4\text{ hexagons}
\]

1a.) Complete the table showing the number of hexagons in a chain and the perimeter (number of outside edges).

<table>
<thead>
<tr>
<th>NUMBER OF HEXAGONS</th>
<th>PERIMETER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
</tr>
</tbody>
</table>

1b.) Describe the process for determining the perimeter for 100 hexagons, without knowing the perimeter for 99 hexagons.

Well... use the equation \( s = 6n - 2(n-1) \), where \( s \) is the total number of sides and \( n \) is the number of hexagons. That is, 6 sides per hexagon minus the overlapping sides which number twice the quantity of hexagons less 1. So

\[
5 = 600 - 2(99) \Rightarrow 600 - 198 = 402 \text{ sides (i.e., perimeter is 402)}
\]

\(^1\)Adapted with permission from Balanced Assessment/MARS (Zawojewski et al., 2000) with permission.
1c.) Write a formula to describe the perimeter for any number of hexagons in the chain (it does not need to be simplified).

\[ \text{perimeter} = \# \text{ of hexagons} - 2(\# \text{ hexagons} - 1) \]

1d.) Explain why you think your formula in (1c) is correct?
Because it is? Because it accurately describes the perimeter - 6 sides per hexagon, subtract 2 for each overlapping join, \( \# \) of overlapping joins is number of hexagons - 1, etc... it makes sense.

1e.) Suppose instead of regular hexagons, you form a chain with regular \( n \)-gons (where \( n \) is any number of sides). Describe the process for determining the perimeter for any number of regular \( n \)-gons in a chain.
Assuming the polygons \( n \)-gons behave just like the hexagons, overlapping on one side, then the formula is the same, just with \( n \) replacing the 6.

1f.) Write a formula to describe the perimeter for any number of regular \( n \)-gons in a chain (it does not need to be simplified).

We'll use \( x \) for \( \# \) of \( n \)-gons now...

\[ \text{perimeter} = n \cdot x - 2(x - 1) \]

1g.) Explain why you think your formula in (1f) is correct?
Why wouldn't it be? The situation seems the same as previously, that is the same number of sides overlap \( 2(x - 1) \). Just different \( \# \) of sides per \( n \)-gon.
APPENDIX B: FUNCTIONS LEARNING PROGRESSION SCORING GUIDE FOR HEXAGON PATTERN ITEM (GP-1)

<table>
<thead>
<tr>
<th>Functions Learning Progression</th>
<th>GP-1 Exemplars</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA – Extended Abstract. Applies knowledge of functions for hexagons to a generalizable formula for n-gons.</td>
<td>Symbolic function for n-gon (1g) is correct and related to the symbolic function for hexagon (1c).</td>
</tr>
<tr>
<td>R – Relational. Relates information from hexagons to determine functions for n-gons.</td>
<td>Verbal function for n-gon (1e/1g) is correct and related to verbal function for hexagon (1b).</td>
</tr>
<tr>
<td>MS – Multi-structural. More than 2 functions (tabular, verbal, symbolic) are clearly connected.</td>
<td>Table (1a), verbal description (1b), and equation (1c) are correct and connected.</td>
</tr>
<tr>
<td>US – Unistructural. 2 functions are connected.</td>
<td>Table (1a) and equation (1c) is connected OR Table (1a) and verbal description of 100 hexagons is connected (1b) (Note: Verbal description must demonstrate knowledge beyond “adding four”).</td>
</tr>
<tr>
<td>PS – Prestructural. 1 function correct.</td>
<td>Table (1a) correct.</td>
</tr>
<tr>
<td>PA – Pre-algebraic. 1 function partially correct.</td>
<td>Table (1a) is partially correct (Note: Student may mention “adding four” as the only substance to the problem).</td>
</tr>
<tr>
<td>OT – Off Target</td>
<td>Nothing correct. Makes no sense.</td>
</tr>
<tr>
<td>DK – Don’t Know</td>
<td>Writes “I don’t know.”</td>
</tr>
<tr>
<td>NR – No Response</td>
<td>Blank</td>
</tr>
</tbody>
</table>

APPENDIX C: MEASUREMENT MODEL

Mathematically, this model is represented as

$$P(X_{sik} = 1 | i, ,') = \frac{\exp(b_{ik}, + a_{ik}')}{{\sum}_{k=1}^{K_i} \exp(b_{ik}, + a_{ik}')},$$

to fit the combination of dichotomous and polytomous responses. This variation of the Rasch model is called the Multidimensional Random Coefficients Multinomial Logit model (MRCML; Adams, Wilson & Wang, 1997) where $P(X_{sik} = 1 | i, ')$ denotes the probability of a correct response and:

- $s$, $i$, and $k$ subscripts represent respondents, items and categories, respectively,
- $\hat{e}$ represents a vector of latent variables (MD if vector is longer than 1),

---

The ConstructMap software, developed by Berkeley Evaluation & Assessment Research (BEAR) Center (Kennedy et al., 2006), was used for calibration.
• \( \mathbf{i} \) represents a vector of item parameters,
• \( \mathbf{b}_{ik} \) represents a scoring vector, and
• \( \mathbf{a}^i_{ik} \) represents a design vector (these are not parameters, but are specified by the user).

### APPENDIX D: EXCERPT FROM R-CRA, INTERPRETED BY TEACHERS SS AND ST

<table>
<thead>
<tr>
<th>Scaled Score</th>
<th>Distribution of Students (n=47)</th>
<th>Estimated Average Proficiency in Math Courses (Lower/Upper Range of 95% Confidence Interval)</th>
<th>STUDENTS ARE ACTIVELY LEARNING...</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td>6th grade (-0.98, -0.60) 7th grade (-1.06, -0.46) 9th grade (-1.32, -0.03)</td>
<td>The SIGNIFICANCE of the connections and how to APPLY them or GENERALIZE.</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>How to MAKE CONNECTIONS across more than one representation.</td>
</tr>
<tr>
<td>1</td>
<td>XXX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>XXXXX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>XXX</td>
<td>8th grade (-34, 26)</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>XXXXXXX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>XXXXXX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td>XX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each X represents 1 student. Dashed line (\) represents average student proficiency.
APPENDIX E: EXCERPT FROM R-CRA INTERPRETED BY TEACHERS BY AND JS

<table>
<thead>
<tr>
<th>Scaled Score</th>
<th>Distribution of Students (n=64)</th>
<th>Estimated Average Proficiency by Math Course (95% Confidence Interval)</th>
<th>STUDENTS ARE ACTIVELY LEARNING...</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each X represents 1 student. Dashed line (-) represents mean proficiency.

The SIGNIFICANCE of the connections and how to APPLY them or GENERALIZE.

How to MAKE CONNECTIONS across more than one representation.

The PREREQUISITE SKILLS to make connections across representations of functions.

- Double Block (-.1, .38)
- Algebra II (-.1, .34)
- Trig Honors (.61, 1.0)