*Research Report*

# *Defining Mathematics Competency in the Service of Cognitively Based Assessment for Grades 6 Through 8*

***Edith Aurora Graf***

***December 2009***

***ETS RR-09-42***

# Defining Mathematics Competency in the Service of Cognitively Based Assessment for Grades 6 Through 8

Edith Aurora Graf
ETS, Princeton, New Jersey

December 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html



**Abstract**

This report makes recommendations for the development of middle-school assessment in mathematics, based on a synthesis of scientific findings in cognitive psychology and mathematics education. The focus is on background research, rather than test specifications or example tasks. Readers interested in early development and pilot efforts associated with the Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL) project in mathematics (for which this review helped provide a theoretical foundation) should consult Graf, Harris, Marquez, Fife, and Redman (2009).The organization of the report is motivated by the evidence-centered design (ECD) approach to assessment developed by Mislevy and colleagues (e.g., see Mislevy, Steinberg, & Almond, 2003). The first section consists of a broad literature review that characterizes mathematical competency with respect to both content and process. Subsequent sections discuss: how to model mathematical competency at the middle school level, the kinds of evidence that reflect the level of student competency and support future learning, and how to design tasks that elicit the target evidence.

Key words: Mathematics assessment, mathematics cognition, mathematical competency, mathematics task design

**Acknowledgments**

Thanks to Randy Bennett, Brent Bridgeman, Jim Fife, Jeff Haberstroh, Liz Marquez, and Val Shute for their reviews and comments on earlier versions of this manuscript. Thanks to Isaac Bejar, Diane Briars, and Drew Gitomer for advice, particularly in referring me to helpful sources. Thanks to Jim Fife and Vicky Pszonka for formatting the equations. Finally, thanks to Waverely VanWinkle and Darlene Rosero for their assistance with formatting. The efforts of these contributors are appreciated and any errors are the sole responsibility of the author.

## Table of Contents

## Guiding Assumptions

The purpose of this report is to make recommendations for the development of middle-school assessment in mathematics, based on a synthesis of scientific findings in cognitive psychology and mathematics education. The focus is on background research, rather than test specifications or example tasks. Readers interested in early development and pilot efforts associated with the Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL) project in mathematics (for which this review helped provide a theoretical foundation) should consult Graf, Harris, Marquez, Fife, and Redman (2009).

The first section consists of a broad literature review that characterizes mathematical competency with respect to both content and process. The following issues are discussed: what is important for students to learn, what students have difficulty learning, and how learning (especially in areas of difficulty) may be facilitated. The focus on learning is deliberate, since the ultimate goal for assessment should be not only to measure competency but to encourage improvement. The first section is an attempt at what is referred to as the domain analysis stage of evidence-centered design (ECD), a principled approach to assessment design developed by Mislevy and colleagues (e.g., Mislevy, Steinberg, & Almond, 2003). Domain analysis includes the background information needed for the development of an evidence-centered design conceptual assessment framework, or CAF.

The CAF consists of three components: a student model, evidence models, and task models (see Mislevy, Steinberg, & Almond, 2003). For each component of the CAF, there is a corresponding section in this document. Section two presents the first draft of a competency model for middle-school mathematics assessment.

Section three provides recommendations and considerations pertaining to the development of evidence models. The competency model can be applied to either formative assessment or accountability assessment, since both may share a common conceptual base. For example, in the CBAL project (see Bennett & Gitomer, 2009) the formative and accountability assessment components share a common competency model, and both components include tasks that are designed to be learning events as well as assessment items. That is, many of the CBAL tasks are extended and have real-world settings that require complex responses. These tasks usually require at least several responses, and some provide opportunities for simulation-based interactions. The CBAL accountability and formative

components also differ in a number of important respects, and these differences will guide how the evidence models for each are developed.

First, they serve different goals. One purpose of the CBAL accountability component is to satisfy the requirements of no child left behind, while providing an alternative to a single, end-of-year assessment. In contrast, the purpose of the CBAL formative component is to provide information to teachers so they can guide instruction on a daily basis, or during the course of a lesson. The evidence from the accountability component may be used for formative purposes, but evidence from the formative component may never be used for accountability purposes. As weaknesses in student understanding are identified from the CBAL accountability results, they may be used to inform the development of the formative component, so that the formative materials directly address areas of student difficulty.

Second, the administration modes for the two systems are very different. Assessments developed for the CBAL accountability system will be administered at multiple time points across the year, as periodic accountability assessments (PAAs). Each PAA will last approximately one class period. In contrast, administration of the formative component will be much more flexible—teachers will use tasks as they see fit.

Third, although there may be many similarities in the tasks and responses from the two assessment systems, the evidence accumulation procedures may differ substantially. For example, evidence from successive PAAs will be accumulated across time points to establish a composite of each student's performance by the end of the year.[1] Evidence from formative assessments, however, may be used to informally assess progress between PAA administrations. The third section of this document concerns issues related to the development of evidence models, and the kinds of evidence that are most important for each of the two assessment systems.

The fourth section outlines task design principles we should follow, as well as requisite features for the tasks. As mentioned earlier, many of the CBAL tasks are extended and include complex response types. Other CBAL tasks are more concise, consisting of a relatively short stem and one or two prompts. While only the extended tasks include simulations or require interactions, either kind of task may require a student to provide one or more of the following: (a) a numeric answer, (b) an expression or equation, (c) a graph, (d) a selection of an option or set of options, or (e) a text response. These design features are

intended to elicit richer evidence of student understanding, but they are also intended to help students learn as they work through the tasks.

## Mathematical Competency: Characterizing Foundations

At the most general level, competency in mathematics is characterized both in terms of content (what mathematics students should know) and process (how students should go about doing and understanding mathematics). This distinction is reflected in the organization of the document *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 2000), which includes both *content standards* and *process standards* for students of mathematics in K-12. The standards are grounded in research and have a long history of development and revision (the original version, *Curriculum and Evaluation Standards for School Mathematics*, was released in 1989).

The content standards include (a) numbers and operations, (b) algebra, (c) geometry and measurement, and (d) data analysis and probability; the process standards include (a) problem solving, (b) reasoning and proof, (c) communication, (d) connections, and (e) representation. All of these standards are central to the study and practice of mathematics, and all have been the subject of research (though different sources use different terminology, and some areas have been much more heavily researched than others). The purpose of this section is to summarize cognitive psychology and mathematics education research findings on both the content and process aspects of mathematical competency.

### *Core Content*

Mathematics curricula vary with respect to the coverage of topics, the sequence in which the topics are taught, and the extent to which mathematics instruction is integrated with instruction in other content areas. Nevertheless, there are common elements that are fundamental to any mathematics curriculum at a given level. This is in part a result of the standards-based reform movement, but it is also implicit in the nature of mathematics as a domain: while there have been trends (and heated debates) over the years regarding which topics should be emphasized, there are certain essential building blocks that students must master before they can meaningfully explore other topics.

Topics in mathematics can be located on branches of a tree (e.g., Hale, 2002). Note that the analogy is made to a botanical tree, not a mathematical tree, which has a precise

definition. The height of a topic on a tree is an indication of its complexity, and its connections with lower branches indicate how the topic subsumes other more basic topics. For example, Hale's tree representation includes more than 20 mathematical topics. Logic, set theory, and number systems occupy locations on the trunk, and algebra, geometry, and analysis each occupy primary branches. Calculus and statistics are located at higher points on the tree, but they are on branches that connect to the topics below. The term *arithmetic* is less often used than it once was—the term *numbers and operations* is used in the NCTM standards (NCTM, 2000), and the term *number properties and operations* is used in the National Assessment of Educational Progress (NAEP) mathematics framework (National Assessment Governing Board, 2007). These classifications include arithmetic, but also refer to knowledge of number systems and the conceptual understanding of number concepts.

The tree can be used as a representation to show how mathematical topics are interconnected and build on each other. Needless to say, there is not perfect agreement among content experts about how topics are interrelated, and in the course of their work mathematicians sometimes discover new connections among branches that were previously considered unrelated, but there is general agreement that numbers and operations, algebra, geometry, measurement, and probability and data analysis are all fundamental topics for K-12 students to learn. These topics are reflected in both the NCTM content standards and the NAEP mathematics assessment framework, as well as in many state standards.

### *Numbers and Operations*

In the report *Adding It Up: Helping Children Learn Mathematics,* The Mathematics Learning Committee from the National Research Council (Kilpatrick, Swafford, & Findell, 2001) summarized a large body of research on mathematics learning in grades K-8 and made recommendations for improving student performance. The committee was charged with focusing on essential skills that were central to continued development of mathematics proficiency in K-8. The committee members recognized that they could not focus on all of the important content areas, so they chose to focus on the concept of number. They provided two strong arguments for this focus: other areas of mathematics build on the concept of number (it inhabits the "trunk" of the mathematical tree), and student learning in this area has been well researched. Number concepts are strongly emphasized in the early and middle grades; there is a shift to algebra later on. In the NAEP 2005 mathematics framework

(National Assessment Governing Board, 2005), 40%, 20%, and 10% of the items assess number properties and operations at grades 4, 8, and 12, respectively. The concept of number includes the study of negative numbers, rational numbers, and proportional reasoning, all of which traditionally pose difficulty for students. It also includes emphasis on approximation and evaluating the reasonableness of results, a keystone in most mathematics standards. As cited in *Adding It Up* (Kilpatrick et al., 2001), Carpenter, Corbitt, Kepner, Lindquist, and Reys found that 55% of 13-year-olds selected either 19 or 21 as the correct response to the following NAEP assessment item: $\frac{12}{13} + \frac{7}{8} = ?$. These responses reflect a lack of understanding about how to add fractions (19 is the sum of the numerators; 21 is the sum of the denominators), but they also suggest that students did not evaluate them for reasonableness (the answer must be less than two).

Most modern versions of national and state mathematics standards emphasize the importance of evaluating the reasonableness of results, and this is part of understanding the concept of number. For example, if a student is asked to calculate the length of a physical object and obtains a negative result, he or she should recognize that either an error has been made or that the problem has been constructed incorrectly. Students are expected to be able to approximate, and to have a sense of quantity and magnitude (sometimes called *number sense*). Tasks that measure proficiency with approximation may explicitly direct the student to approximate, or, approximation may be used as a supporting strategy (for example, as in the NAEP item above).

Working with rational numbers can be particularly difficult because they occur in a number of different forms, each with different notations and different interpretations. Students must appreciate the meaning of fractions, decimal fractions, and percents and be comfortable converting among them. When finding the sum of two fractions with different denominators, many students will add the corresponding numerators and denominators, bypassing the step of finding a common denominator (Siegler, 2003). Silver (as cited in Siegler, 2003) found that adults taking community college mathematics courses make the same error. It has been suggested that errors like this may occur because students have difficulty perceiving a fraction as a single quantity, operating on it instead as if it were two distinct numbers (Kilpatrick et al., 2001, p. 235).

Resnick et al. (as cited in Siegler, 2003) found that students sometimes judge that decimal fractions with more digits are always larger than decimal fractions with fewer digits. This too has been interpreted as a possible result of students' experiences from working with whole numbers, where a number that has more digits is also larger. This suggests that while an understanding of the whole number system is certainly a prerequisite for learning about rational numbers, it can also interfere with the learning of rational numbers.

Understanding the concept of a ratio and how to operate with ratios is already difficult; interpreting a proportion (which is an equality between ratios) additionally requires an understanding of the notion of equivalence, which many students lack (e.g., Kieran, 1992). Using additive operations inappropriately in multiplicative contexts is a very common error among students; in proportional reasoning, this error is often referred to as the *incorrect addition strategy* (Hart, 1984). This misunderstanding is pervasive and crosses mathematical content areas (e.g., Karplus, Pulos, & Stage, 1983; Noelting, 1980; Vergnaud, 1983). A student might apply it in a strictly numeric context (e.g., $\frac{2}{7} = \frac{3}{8}$) or in a geometric context. For example, suppose a student is given a right triangle with sides of lengths 3, 4, and 5 units and a similar triangle where the shortest side is of length 9 units. When asked for the lengths of the other two sides of the similar triangle, a student who uses the incorrect addition strategy will respond that the other two sides are lengths 10 and 11 units. Apparently many students have difficulty in making the leap from additive to multiplicative models. In general, reasoning about ratios and proportions may be difficult because it involves reasoning about relationships between quantities. Many eighth-graders and adults had difficulty with ratio and proportion tasks that require reasoning rather than computation (Lesh & Lamon, 1992, p. 29), suggesting that understanding ratios and proportions (beyond performing basic computations with them) is a conceptual leap that is not necessarily made by adulthood.

In sum, a strong case can be made for focusing on the concept of number and number systems as a foundational content area. It occupies the "trunk" of the mathematical tree and is well represented in the curriculum in the early and middle grades. Rational numbers, negative numbers, and proportional reasoning are all well-researched and encompassed by this content area.

### *Pre-Algebra and Algebra*

Traditionally, Algebra 1 is most often taught in the ninth grade, but in many places there is a push to teach it earlier, in eighth grade. Algebra is emphasized later in the curriculum than numbers and operations, though the two topics overlap. Algebra can be taught as a single course in a single grade or as a strand that is addressed across grades, and it can be integrated to a greater or lesser extent with other content areas like geometry (National Research Council, 2000, p. 145). In the NAEP 2005 mathematics framework (National Assessment Governing Board, 2005), algebra is assessed by 15%, 30%, and 35% of the items in grades 4, 8, and 12, respectively. There are algebra standards across grades in the NCTM standards as well, beginning with the K-2 grade band (NCTM, 2000). This is more consistent with the view that students should learn algebra, or at least algebra concepts, over a long period of time. Competency in algebra comprises at least two main components: algebraic manipulation and algebraic representation. Problem solving typically involves use of both components of competency, and each is necessary for understanding algebraic concepts. While the committee that prepared the *Adding It Up* report selected the concept of number as a content focus, the RAND mathematics study panel (RAND Mathematics Study Panel & Ball, 2003) recommended that algebra should be the first content area to receive focus, for many of the same reasons: algebra is foundational, and without a solid understanding of it, most mathematics courses that are more advanced are inaccessible.

*Algebraic manipulation.* Tasks that assess facility with algebraic manipulation often provide the student with an algebraic expression or equation and ask the student to operate on it (for example, the student may be asked to simplify an expression or to solve for a variable in an equation). Such tasks are sometimes entirely procedural, but conceptual understanding often facilitates their solution. For example, consider the following hypothetical item:

$$\text{If } 3 - 9x = 4, \text{ find } 3x - 1.$$

This item can be solved by first solving for $x$ and then finding the quantity $3x - 1$. But because this involves a lot of computation, it is more time-consuming and increases the chance for error. Alternatively, if the student recognizes that $3 - 9x = -3(3x - 1)$, the item can be solved quickly by dividing 4 by −3. So in this case the solution to an apparently procedural item is greatly facilitated by insight. Insight (or perhaps schema recognition) is

also involved in finding the roots of a "disguised" quadratic equation (an equation that is not quadratic with respect to its original variable, but that may be expressed in quadratic form when an appropriate substitution is made), or in making an appropriate substitution to evaluate an integral.

It is probably apparent from this discussion that what an item assesses depends on how it is solved. I will not discuss this point at length here (this is reserved for the section on strategies), but it demonstrates that conceptual insight can facilitate the solution of tasks that appear to be primarily procedural. Similarly, while procedural errors sometimes demonstrate only a lack of procedural knowledge, they may demonstrate gaps in conceptual understanding. The relationship between conceptual understanding and procedural fluency was emphasized in *Adding It Up* (Kilpatrick et al., 2001). The following excerpt from Harel (as cited by Kaput, 1999, pp. 140-141) illustrates how errors in a primarily procedural task may reflect conceptual misunderstanding:

> The high school student in this example was attempting to solve the inequality $((x-1)^2 > 1)$ . When asked to explain how she arrived at $x > 1$, she responded that "The solution to the equation $(x-1)(x-1) = 0$ is $x = 1$, $x = 1$." She then crossed out the three equality signs and above each wrote an inequality sign >, noting that "$x$ is greater than 1." When she was then asked to solve $(x-1)(x-1) = 3$, she wrote: "$(x-1) = 3$, $(x-1) = 3$."

Harel observed that the student was apparently attending to the surface features of the problem, which are extremely similar to the surface features present in a quadratic equation. She then followed the procedure for finding the roots of a quadratic equation, without attending to the meaning of the inequality. In the last example, the student again attended to surface features, apparently not realizing that $(x-1)(x-1) = 3$ does not imply one of the two factors must equal 3.

Errors like the ones above are possible to make in haste and, no doubt, are sometimes due to a transient slip. Protocol studies where students have been interviewed, however, suggest that many such errors are due to real misunderstanding. Lee and Wheeler (1989) presented students with several algebraic statements and asked them to determine whether a given statement was definitely true, possibly true, or never true—students were also asked to justify the response. One of these statements was as follows:

$(a^2 + b^2)^3 = a^6 + b^6$ (Lee & Wheeler, 1989, p. 42)

Half of the 10th-grade students queried believed this statement was true; the following was among the justifications that were provided:

> This statement is definitely true. There are several laws in dealing with exponents. And the one that applies here is you multiply the number (outside the bracket) with those exponents inside the bracket. You don't add them like you normally do. If you had an example like $a^2 + a^3$ you add them so you get $a^5$ but the brackets tell us to multiply. (Lee & Wheeler, 1989, p. 42)

While some students responded with "it's a rule" as their justification, a number of them provided explanations similar to that above. This example suggests that algebraic errors are not always the result of a transient slip. Note, however, that this does not imply that misunderstanding about a concept always results in the same error or even in an error at all. Students have been observed to apply a large number of algebra *mal-rules*, or incorrect rules, in solving algebraic manipulation items (e.g., Payne & Squibb, 1990; Resnick, Cauzinille-Marmeche, & Mathieu, 1987; Sleeman, 1984). These are idiosyncratically applied and their incidence seems to vary across populations; however, manipulations where parentheses are involved do appear to consistently pose difficulty.

*Algebraic representation*. Algebraic representation is concerned with constructing models that describe situations in mathematical terms. This definition is broader than what was traditionally referred to as algebraic representation, which usually meant representing a word problem in symbolic form, but it is also more consistent with current mathematics standards. As a result of standards-based reform, mathematical proficiency is now characterized much more broadly in terms of what students should be able to do in mathematics (see Schoenfeld, 2006, for a discussion). The expansion in requirements is most evident in terms of expectations for students' fluency with alternate representations, or models. Models often include any of the following: expressions, equations, diagrams, tables, or graphs. Lesh and Lamon (1992) argued that model representation is essential to capturing large amounts of information in concise, operable form. Algebra, and the study of functions in particular, lends itself to the use of alternate models, and this is reflected in mathematics standards. The following quote (NCTM, 2000, p. 38) suggests that while facility with

alternate representations is expected, using alternate representations may also help students learn the concept of function more completely:

> Many college students understand the notion of function only as a rule or formula such as "given $n$, find $2^n$ for $n = 0, 1, 2$, and 3" (Vinner & Dreyfus, 1989). By the middle grades, students should be able to understand the relationships among tables, graphs, and symbols and to judge the advantages and disadvantages of each way of representing relationships for particular purposes. As they work with multiple representations of functions—including numeric, graphic, and symbolic—they will develop a more comprehensive understanding of functions (see Leinhardt, Zaslavsky, & Stein, 1990; Moschkovich, Schoenfeld, & Arcavi, 1993; NRC, 1998).

In other words, working with tasks that require the use of multiple representations may play a formative role in algebra instruction.

Translating statements to algebraic expressions has a long history of causing difficulty for students. The famous "Students and Professors" statement (Clement, Lochhead, & Monk, 1981, p. 288; Clement, Lochhead, & Soloway, 1979, Table 1) is as follows:

> Write an equation using the variables $S$ and $P$ to represent the following statement: "There are six times as many students as professors at this University." Use $S$ for the number of students and $P$ for the number of professors.

Among college students, a common incorrect response to this item is "$6S = P$" (this is known as the *variable reversal error*). In a study of 150 engineering students, 37% of 150 freshman engineers responded incorrectly to this item; two thirds of the incorrect responses were incorrect due to the variable reversal error (Clement, Lochhead, & Monk, 1981). The error rate for nonscience majors was higher; the statement was translated incorrectly by 57% of 47 nonscience majors. These error rates are pretty typical of the rates in replicated studies.

In the context of studies on arithmetic word problems, Lewis and Mayer (1987) distinguished between items with *consistent language* versus items with *inconsistent language*. Items that use consistent language suggest operations that are consistent with the correct representation; items that use inconsistent language suggest operations that are inconsistent with the correct representation. Although the students and professors statement is not an arithmetic item, it uses inconsistent language and is much more difficult than

algebraic translation items that use consistent language (e.g., Graf, Bassok, Hunt, & Minstrell, 2004).

### *Making the Connection Between Numbers and Algebra*

So far in the discussion of content, I have described arguments for focusing on numbers and operations on the one hand, and arguments for focusing on algebra on the other. Topics in mathematics are connected, however, and these connections deserve attention in the context of assessment, as well as in the context of instruction. But there is evidence that students do not perceive the connections between mathematical topics, particularly where connections between arithmetic and algebra are concerned.

Lee and Wheeler (1989) presented 10th-grade students with two types of tasks. One set of tasks consisted of algebraic statements, and each student had to explain whether one of the statements was definitely true, possibly true, or never true (one of these statements, $(a^2 + b^2)^3 = a^6 + b^6$, was discussed in the preceding section). Lee and Wheeler noted that although it could be shown through the use of numeric counterexamples that none of the statements were always true, only 10 out of 268 students made an attempt to substitute numbers into a statement. Students who justified $(a^2 + b^2)^3 = a^6 + b^6$ as adhering to a rule were asked to substitute values into the equation. Some of the students were not surprised by the resulting contradiction, and when asked about it, suggested that they wouldn't necessarily have expected algebra and arithmetic to produce consistent results.

Lee and Wheeler (1989) designed another set of tasks to suggest numeric representations—variables were not used. One of these tasks was as follows:

> A girl multiples a number by 5 and then adds 12. She then subtracts her starting number and divides the result by 4. She notices that the answer she gets is 3 more than the number she started with. She says, "I think that would happen, whatever number I started with." Is she right? Explain carefully why your answer is right. (p. 47)

Although the tasks in this set could be justified using algebra, the majority of students provided demonstrations by using a finite set of numeric examples—which is insufficient. For the question above, algebraic approaches were particularly rare: only 9 out of 118 students attempted an algebraic justification. Lee and Wheeler (1989) interpreted their findings to suggest that, for many students, arithmetic and algebra are dissociated branches of

mathematics characterized by distinct sets of procedures. Borchert (2003) found that even among college students, there is a "dissociation" between arithmetic and algebra.

A similar finding is described in the work of Resnick et al. (1987). In a set of interviews, they asked children between the ages of 11 and 14 to determine whether or not pairs of expressions were equivalent. Each expression pair was shown first with variables, and then with numbers. They observed that the children used one of three strategies for judging equivalence. Resnick et al. referred to these alternate strategies as *calculation*, *rule-based evaluation*, and *approximate evaluation*. Calculation involved calculating the value of each expression (by substituting numbers if necessary) to determine whether there was a match. Rule-based evaluation involved applying algebraic rules to determine equivalence (sometimes these were rules the children had been taught, other times they were incorrect rules the children had invented). Approximate evaluation was a form of analysis in which the children did not consider the specific quantities involved; rather, they considered whether one expression was greater than, less than, or equal to the other expression. Resnick et al. noted that:

> For the most part, these different strategies functioned as 'islands of knowledge' (cf. Lawler 1981), communicating very little with each other. This meant that children rarely used knowledge of one type to constrain or justify judgments of another type. On the other hand, one of the strategies would sometimes intrude on another to produce errors. (p. 179)

Again, this suggests that many students do not perceive a connection between arithmetic calculation and the application of algebraic rules. More than that, it suggests that partial knowledge of the two systems can create conflicts during execution. In one example from Resnick et al. (1987), a student correctly removed the parentheses from the expression $14 – (9 + 3)$ to yield $14 – 9 – 3$, but then subtracted 6 from 14 to yield 8. In other words, the student calculated $14 – (9 – 3)$, even though the parentheses were already removed (p.180). Resnick et al.'s interpretation was that in this case the arithmetic error was due to a misapplication, or intrusion, of a formal rule for algebraic manipulation.

Greeno et al. (1986) noted differences in the cognitive demands between arithmetic and algebra, and suggested that these differences might partially account for beginning students' fragmentary knowledge of algebra. They made the following observation:

> When we considered the cognitive requirements of elementary algebra problems, we realized that they have a fundamentally different structure from that of almost all the tasks students learn to perform in arithmetic. In arithmetic, almost all problems involve *evaluating* symbolic expressions, but in algebra, most problems involve *transforming* symbolic expressions into equivalent expressions. The operators and goals for transformation tasks differ significantly from those of evaluation tasks. (p. 34)

Obviously arithmetic and algebra share much in common, including operators and relations. But as pointed out in the quote above, the goals for arithmetic and algebra are usually quite different—arithmetic tends to involve evaluation while algebra tends to involve transformation. From a cognitive perspective, these goals are very different—the latter is somewhat more "open-ended" in the sense that transformations tend to be less prescribed than evaluations.

Even though students often do not spontaneously perceive the relationship between arithmetic and algebra, they can be encouraged to do so. When the appropriate scaffolding is provided, arithmetic and algebra can be mutually supportive rather than conflicting. There are several approaches that can help students make the connection, and all of them involve leveraging their familiarity and facility with arithmetic. For example, in spite of their difficulties with translating inconsistent relational statements, college students generally have no difficulty solving arithmetic problems such as the following: "There are 3450 students. If there are 6 times as many students as professors, how many professors are there?"—proportion correct was 0.92 (Martin & Bassok, 2005). Borchert (2000) found that giving students such an arithmetic problem prior to translating an inconsistent relational statement improved performance. Similarly, Bernardo and Okagaki (1994) found that providing students with either symbolic information (e.g., a reminder about the definition of a variable, and that it can assume different values) or arithmetic problem context prior to the translation task improved their performance on translating inconsistent statements to equations.

Koedinger and Anderson (1998) found that students who solved arithmetic problems prior to formulating a corresponding algebraic expression showed greater pretest to posttest gains on a test consisting of both arithmetic and expression items than students who formulated an algebraic expression prior to solving corresponding arithmetic problems.

Finally, Wollman (1983) found that students improved performance with constructing equations when they checked them by substituting the variables with numeric values.

In combination, these results suggest that (a) students do not spontaneously make the connection between variables and values, and (b) reminding students about the relationship between values and variables can improve performance. It should be noted, though, that these attempts can be short-lived. Rosnick and Clement (1980) found that even though students could learn to generate equations correctly, they often lacked conceptual understanding of their equations. Replicating earlier results, Graf, Bassok et al. (2004) found that students who answered a related word problem and explained the solution prior to translating an inconsistent statement to an algebraic equation had much higher equation performance relative to a control group. On a transfer test, however, students who had solved word problems did not do any better at translating inconsistent statements to equations than students from the control group. However, the interventions in Rosnick and Clement and Graf et al. were both short in duration; it is reasonable that an error as pervasive as the variable reversal error would have to be remediated over the long term.

The preceding discussion suggests that encouraging students to perceive the correspondence between arithmetic and algebra may be an effective instructional approach, especially if students can learn to do it spontaneously. Resnick et al. (1987) made the case for relating formal algebra to concrete situations. They found that some students can create stories that support their understanding of algebraic rules but that other students struggle with this task. But they argued that students' skill with relating formal rules and situations should be developed, because understanding the links between algebraic rules and situations is necessary in order to represent new situations as mathematical models. Also, they argued that if students understand how algebraic rules and situations interact, algebraic knowledge and situational knowledge can be mutually reinforcing, since both provide information directly relevant to problem solving.

Kieran (as cited in Kieran, 1992) designed exercises to enhance students' understanding of algebraic equations and equivalence. She first gave students practice with constructing what she referred to as *arithmetic identities* (sometimes referred to as number sentences, these are equations with numbers and no variables). Students practiced these until they could construct identities using multiple operators on each side of the equals sign. She

then gradually introduced the notion of variables by covering numbers in the identities, Eventually, numbers were replaced with boxes, and finally letters. This kind of approach makes sense, because students can leverage their earlier knowledge of numbers in helping them to understand algebra. Kieran's approach involves designing arithmetic tasks that are something like algebra tasks. Using the terminology from the quote from Greeno et al. (1986), some of these arithmetic tasks were designed to encourage *transformation* as well as *evaluation*, providing a link to algebra but in the familiar domain of arithmetic. This is the approach taken in the *Algebridge* series developed by ETS and the College Board.

Earlier it was mentioned that a more modern interpretation of algebra proficiency involves flexibility with using multiple representations. In this interpretation, knowing how to translate a relational statement to an algebraic equation is not sufficient. A student should also be able to represent the situation in a table or a graph. The student might also be asked to draw pictures (using different icons to represent students and professors, for example).

***Geometry and Measurement***

It has been argued that geometric and measurement models provide students with a way to represent and visualize problems in other areas of mathematics and in real life (NCTM, 2000). Arrays can be used to help students develop what is referred to as the area model of multiplication (Kilpatrick et al., 2001, p. 74). Students can use manipulatives to demonstrate that the product of 5 linear units times 6 linear units is 30 square units. This representation can also be extended to algebra. In the NCTM *Principles and Standards*, it is shown how the area model may be used to help students visualize the binomial expansion, $(a + b)^2 = a^2 + 2ab + b^2$ (NCTM, 2000, p. 238). Similarly, the Cartesian coordinate system can be used to solve problems in algebra; students can solve a system of two equations graphically by plotting the two lines and finding the intersection point.

Because geometry allows students to reason with a minimum of symbols, it often provides their first exposure to mathematical argument and proof. For example, Bastable and Schifter (as cited in Kaput, 1999) found that students in a third-grade class were able to use the area model of multiplication (described above) to demonstrate that whole numbers are commutative under multiplication. The students did this by showing that any *m* by *n* array could be rotated 90 degrees to produce an *n* by *m* array. At the middle-school level, students can use geometric constructions to prove the Pythagorean theorem. There are a number of

interesting proofs by dissection for the Pythagorean theorem (Weisstein, 2006); these could be explored through the use of cutouts as concrete manipulatives. Many Web sites also offer online manipulatives for students to explore. For example, the National Library of Virtual Manipulatives provides two puzzles for students to explore the Pythagorean theorem (National Library of Virtual Manipulatives, 1999). Tools like Geometer's Sketchpad also allow students to learn about geometric relationships in an exploratory fashion.

Tatsuoka, Corter, and Tatsuoka (2004) and Birenbaum, Tatsuoka, and Yamada (2004) used the rule space method to develop attribute profiles for 20 different countries, based on performance data from the 1999 Third International Mathematics and Science Study–Repeat (TIMSS-R) mathematics items. Among the findings they noted was that, relative to other countries, U.S. students had particular difficulty with geometry (the associated mastery probability for the geometry attribute was low). A principal components analysis showed that geometry, logical reasoning, proportional reasoning, and higher order thinking skills all loaded on the same component. Tatsuoka et al. speculated that this might not be coincidence—since geometry is sometimes used to introduce proof, perhaps students with experience in geometry might also develop their higher order thinking skills. They also found it surprising that while higher order thinking skills were highly correlated with geometry skill, they were not highly correlated with algebra skill. They suggested that it might actually be better to teach mathematical thinking skills through geometry than algebra (p. 920).

The suggestion made by Tatsuoka et al. (2004) that geometry might be better than algebra for supporting the development of mathematical thinking is tentative. First, the conclusions drawn may only be considered with respect to the particular algebra and geometry items used in the TIMSS-R. Nevertheless, it is an interesting finding and the efficacy of using geometry to cultivate students' higher order thinking skills would be a valuable future line of research. The finding that proportional reasoning and geometry skill were correlated is also interesting—students use proportional reasoning when working with similar figures, and the incorrect addition strategy discussed earlier has been found to be more common in geometric contexts (Kaput & West, 1994).

Measurement is also important for students to master because of its close connections with science and everyday life (NCTM, 2000). As they learn measurement, students also

develop an understanding of unit and scale, which affords another way of thinking about proportional reasoning.

### *Probability, Statistics, and Data Analysis*

Much of what we know about the way people reason in statistics comes from the judgment and decision-making literature and began with the classic work of Kahneman and Tversky. When making predictions, people often behave in accordance with the *representativeness heuristic* and disregard issues such as sample size, base rate information, the regression principle, and some logical rules (Kahneman & Tversky, 1972, 1973, 1982; Tversky & Kahneman, 1971). In general, people tend to predict the likelihood of an event based on perceived similarities between the sample and the population (Tversky & Kahneman). In one example from Tversky and Kahneman, participants were asked to consider two hospitals; at one hospital, 45 babies are born each day, and at the other hospital, 15 babies are born each day. When asked which hospital reports more days on which more than 60% of the babies are born male (assuming that 50% of babies in the population are boys), most participants responded that both hospitals report an equal number of such days; in other words, they neglected the sample size in making their response. Disregard for sample size is pervasive, and even experts with a high level of statistical training sometimes fail to consider it (Tversky & Kahneman).

Another example of people's susceptibility to the representativeness heuristic involves the *conjunction effect* (Kahneman & Tversky, 1982). In this study, participants were given a description of Linda, who had been a philosophy major concerned with human rights issues. Participants were then asked to judge whether it was more likely that Linda was a bank teller or a feminist bank teller. Out of a large sample of undergraduates without statistical training, 86% responded that it was more likely that Linda was a feminist bank teller.

While Kahneman and Tversky have focused on errors in statistical reasoning, others have focused on examples of rational statistical reasoning and on how statistical reasoning can be instructed (Fong, Krantz, & Nisbett, 1993; Lehman, Lempert, & Nisbett, 1993; Nisbett, Fong, Lehman, & Cheng, 1993; Nisbett, Krantz, Jepson, & Kunda, 1993). Nisbett, Krantz, et al. argued that people do reason statistically under certain conditions and the three factors that influence their reasoning include: "…clarity of the sample space and sampling

process, recognition of the role of chance in producing events, and cultural prescriptions to think statistically…" (p. 27).

For example, Nisbett, Krantz, et al. (1993) found that students do take sample size into consideration when they have reason to believe that the population is heterogeneous, but not when they have reason to believe that it is homogeneous. Their point was that when the population is believed to be homogeneous, operating according to the representativeness heuristic is reasonable thinking rather than fallacious reasoning. Further, even though experts sometimes do not apply statistical reasoning, statistical reasoning is amenable to instruction. Fong et al. (1993) found that students gave statistically based responses more frequently and of better quality when they were instructed in both formal rules and given examples. Lehman et al. (1993) found that graduate students in psychology and medicine showed improved statistical reasoning after two years of graduate school, while students in law and chemistry did not (the GRE® and/or Law School Admission Test scores for the different disciplines were comparable).

Although the work described thus far has focused on undergraduates' difficulties with statistical reasoning, it is likely that they could be addressed earlier in instruction. Statistical reasoning in particular lends itself to simulation activities.

So far in this discussion we have focused on statistical reasoning, but data interpretation, analysis, and display deserve equal attention. Lehrer and Schauble (2000) examined how young children developed and revised data models. In general, they observed that children's data displays were initially not very informative but evolved to be more meaningful through iterative rounds of guided class discussion and subsequent modification of the display. For example, one teacher asked her first-grade students the question: "How do we wake up in the morning?" Each child wrote a response on a sticky note which was stuck to a large poster. After the teacher observed that the display was not easy to read, the children grouped the sticky notes into labeled columns (i.e., Alarm Clock, Mom/Dad, Radio). Further revisions (suggested by the children, with guidance from the teacher) included: organizing the groups into rows instead of columns; creating a grid of equally sized boxes, with one sticky note placed in each box; and numbering the columns. The end result was the equivalent of a horizontal bar graph.

In another exercise from Lehrer and Schauble (2000), each student in a class of third graders was asked to record the number of objects recycled each week for a month in his or her home. The number of objects for each week from each student was written on a sticky note, and the sticky notes were placed on a board. Students were then asked to organize the sticky notes so that they could show the number of recycled objects in a given week. As part of this exercise, students developed distributions based on different groupings, as well as different measures of central tendency. These examples from Lehrer and Schauble illustrate that, with appropriate guidance, young children can begin to think about how to organize, structure, and display data. This certainly implies that more could be done with data organization and display in the middle grades as well. Data organization and display is another area that lends itself particularly well to computer-based exploration. For example, TinkerPlots is a software tool designed to allow students in grades 4-8 to visualize and explore data.

***Summary***

In this section, research findings with implications for instruction and assessment were described, in each of the central mathematical content areas. Mathematical content areas are interconnected rather than distinct, and for this reason can be represented as branches on a tree. The preceding discussion included some examples of how the different branches connect when applied to real mathematics problems. We could safely choose any of the above content areas as a starting point and make meaningful progress in designing a cognitively based assessment system. We could choose numbers and operations because it is arguably the most foundational and has a long history of research behind it. We could focus on algebra because it formalizes mathematical argument, is heavily emphasized in the mathematics curriculum, and affords opportunities for working with multiple models and representations. We could target geometry and measurement for almost the opposite reason: it has been underemphasized in the U.S. curriculum and U.S. students are having particular difficulty with it as reflected by the TIMSS-R results, and it can be used as part of a more informal introduction to proof. Finally, probably in response to workplace demands, statistics and data analysis have recently received greater emphasis in the standards and in the curriculum, so a meaningful contribution could be made here as well.

All of these content areas are important, and since all are addressed by state standards, all need to be represented in any accountability assessment that is developed. Also, since the different content areas are interconnected, we should not include one or two content areas to the exclusion of the others. It is more a question of where to focus. In the middle grades, numbers and operations are central early on, with an increasing emphasis on algebra as students advance towards eighth grade. As discussed earlier, many students do not perceive a relationship between arithmetic and algebra. The literature discussed spans the middle grades through college, and it appears that a general problem at all of these levels is that students have difficulty grasping relationships between numbers and operations and algebra, and that algebraic transformations may be perceived as a set of arbitrary procedures to be memorized. It may be that misunderstandings about algebra that begin in the middle grades will persist through high school and college if they are not addressed. There is evidence, however, that students can leverage their knowledge of number to improve their understanding of algebra. This was the rationale behind the development of *Algebridge*, developed by ETS and the College Board.

For our initial efforts, I recommend that we *focus* on the link between numbers and operations and algebra, but not to the exclusion of the other content areas. For example, Vennebush, Marquez, and Larsen (2005) presented a number of tasks that are primarily identified with other content areas (for example, geometry or data analysis) that lend themselves to algebraic thinking. They also demonstrated how tasks from other content areas may be modified to assess algebra to a greater extent. The assessments we develop should encourage students to recognize and use alternate representations (including expressions, words, graphs, diagrams, and tables) and to understand the relationships between numbers and algebra, as well as other content areas. The focus I am proposing is broader than traditionally conceived notions of arithmetic and algebra, but is more consistent with the thinking behind current standards. The importance of representation to current practice in mathematics assessment is described further in the *Key Processes* section.

### *Key Processes*

While the earlier section focused on the content that is central to K-12 mathematics, this section focuses on what it means to do mathematics. Hoffman and Steen (as cited in Schoenfeld, 1994) describe mathematics as a science of patterns. Consistent with the

mathematics as a science of patterns view, Lesh and Lamon (1992) provided the following definitions for pure and applied mathematics:

- Doing "pure" mathematics means investigating patterns for their own sake, by constructing them and transforming them in structurally interesting ways, and by studying their structural properties.
- Doing "applied" mathematics means using patterns as models (or structural metaphors, or quantitative structures) to describe, explain, predict, or control other systems—with refinements, extensions, and adaptations being made to these models when necessary. (p. 25)

Schoenfeld (1994) emphasized the empirical aspect of the science as patterns definition and used it as a springboard to discuss the experimental quality of engaging in mathematics. Among the points he made were that doing mathematics means collaborating with members of a larger community, and that while mathematical results may be concise and elegant, the process of getting to results is not preordained and may involve many detours along the way.

These perspectives on doing mathematics are reflected in the current *Principles and Standards* (NCTM, 2000). As mentioned earlier, there are five process standards, as follows: (a) problem solving, (b) reasoning and proof, (c) communication, (d) connections, and (e) representation. These are the skills that a student is expected to develop over the course of mathematics instruction. NCTM (2000) defines problem solving as "…engaging in a task for which the solution method is not known in advance" (p. 52). This is a broad definition, but the term *problem solving* is sometimes used more specifically to refer to solving word problems—in this review, the broader definition is assumed. Reasoning and proof refer to the logical processes that are used to develop arguments in mathematics, and communication refers to the expectation that students will develop the capacity to explain and justify their mathematical arguments to others, in speech and in writing. It is expected that their explanations should be clearly articulated and increasingly formal. The connections standard specifies the expectation that students will perceive links between related mathematical ideas. From our earlier discussion, it should be clear that the links are there, but perceiving them is an effortful process. A cognitive interpretation of the connections standard is that it requires students to perceive relationships so that they can adapt strategies and transfer their learning

in the solution of novel tasks. To meet the representation standard, students are expected to use mathematical representations (e.g., graphs, symbols, diagrams) to convey and model mathematical concepts; they are also expected to translate among equivalent representations.

Five interwoven *strands* of mathematical proficiency are discussed in *Adding It Up* (Kilpatrick et al., 2001). They also address process, but while the NCTM process standards focus on activities, these focus on general capabilities and include: (a) *conceptual understanding*, (b) *procedural fluency*, (c) *strategic competence*, (d) *adaptive reasoning*, and (e) *productive disposition*. The first two strands are probably familiar and are also highlighted in the NAEP 1990-2003 mathematics framework.

Conceptual understanding requires that a student appreciate the significance of mathematical principles and recognize how to apply them in various contexts. Procedural fluency requires that a student be facile and efficient with mathematical computations and algorithms. Although it is useful to distinguish between these two strands of proficiency, the following point from *Adding It Up* (Kilpatrick et al., 2001) is important to emphasize: although some tasks may primarily focus on either conceptual understanding or procedural fluency, the strands are most often applied in combination, and are mutually reinforcing—developing procedural fluency can enhance conceptual understanding, and conceptual understanding can lead to greater accuracy and efficiency in the execution of procedures. There is sometimes the perception that procedural fluency and conceptual understanding compete for coverage in the classroom, each at the expense of the other, but ". . . pitting skill against understanding creates a false dichotomy" (p. 122). Sfard (1991) made this point in her framework for how mathematical ideas develop. Her premise was that mathematical notions can be interpreted both operationally (as processes) and structurally (as objects). She argued that operational and structural interpretations are complementary rather than conflicting, and in fact constitute a *duality* rather than a *dichotomy* (p. 9).

Strategic competence refers to the degree to which a student has cultivated successful habits for solving problems in mathematics (Kilpatrick et al., 2001, pp. 124-129). This includes how well students can interpret a problem, how well they can represent it, and how well they can execute a plan to find an answer. Mayer (1983) used the term *strategic knowledge* and defined it as follows: "*Strategic knowledge*–techniques for how to use the various types of available knowledge in solving a given problem, such as setting subgoals"

(p. 354). Adaptive reasoning is used to modify procedures, justify their execution, and verify results (Kilpatrick et al., pp. 129-130). Sometimes in the course of solving a problem, the student will arrive at an impasse, because a particular strategy is not yielding progress towards the ultimate goal. On other occasions, the student may generate a contradiction or unexpected result, suggesting that the selected strategy is in error. In these situations, a student must adapt his or her reasoning to continue making progress or to correct an error. Finally, productive disposition refers to a student's belief that mathematics is a worthwhile and meaningful enterprise, and that mathematical problems are solved through diligent application of effort and concentration.

In the remainder of this section, I discuss three processes that are central to mathematical competency: problem solving, modeling and representation, and argument and justification. Although the NCTM process strands and the strands of proficiency from *Adding It Up* are all important, they do overlap to a substantial degree—the relationships among the different characterizations of competency are included in the following discussion. Also, it is not clear that all of them (e.g., productive disposition) are possible or appropriate to measure *in isolation* as part of a mathematics assessment, although a productive disposition is certainly a requirement for solving a difficult problem successfully. For these reasons, the review is structured with respect to these three processes. Note that as with the NCTM standards, the focus is on activities, since in the design of an assessment this leads naturally to the development of tasks.

### *Problem Solving*

In his classic book, *How to Solve It: A New Aspect of Mathematical Method*, Polya (1957) described four phases for principled problem solving. The first phase is to understand the problem—this includes identifying the given information and specifying the unknowns in the problem. The second phase is to formulate a plan for solving the problem. During this phase, the student should attempt to draw on prior knowledge. If the student knows how to solve a similar problem, he or she may apply the insights and methods used to solve that problem to the new problem. In the third phase, the student executes the plan developed in the second phase. Polya argued that it is especially important that the student justify and check each step during this phase. The fourth and final phase involves reflecting on the solution. This includes verifying the result and consolidating what has been learned so that it

can be generalized to the solutions of new problems. Polya's phases were not intended to characterize the way most students solve problems—rather, they were intended to provide pedagogical support, so that students could solve problems more successfully.

Polya's phases for problem solving have stood the test of time in mathematics pedagogy and are reflected in modern mathematics standards, as well as more recent research in cognitive psychology and mathematics education (Nickerson, Perkins, & Smith, 1985). Analogical reasoning, schema theory, and metacognition are all embodied in Polya's phases. Problem solving is sometimes distinguished apart from conceptual understanding and procedural fluency, but this too is an artificial separation, at least if problem solving is approached in a planful way. Successful execution of Polya's phases requires the use of all five strands of mathematical proficiency as described in *Adding It Up* (Kilpatrick et al., 2001).

Polya's first phase, which involves understanding the problem and framing it in mathematical terms, is a difficult phase for many students, partly due to misconceptions about the nature of mathematics. According to Schoenfeld (1994), "Many if not most students see mathematics word problems simply as cover stories that give rise to computations" (p. 57). Fuson, Kalchman, and Bransford (as cited in Donovan & Bransford, 2005) described a number of common student preconceptions about mathematics. Number one on the list is "Mathematics is about learning to compute" (p. 220). It is a common phenomenon for students to bypass Polya's first stage entirely and rush directly into computation. Doing so can result in negative consequences, however, including inefficient strategies or responses that do not address the question. This is not to say that computation is not important to mathematics—it is essential. But ideally the student should try to understand a problem before attempting computation. So Polya's first phase requires conceptual understanding. One of my mathematics professors appreciated the importance of this first phase and encouraged students to engage in it by first asking if the question was clear. He would then call on a student to state the goal of the problem in his or her own words. Depending on whether or not the question was clearly phrased, discussion among members of the class would ensue. Only when he was convinced that the students in the class had fully understood the nature of the question, the given information, and the unknowns, would he proceed to the next phase.

During Polya's second phase, the student should attempt to draw on prior knowledge. This often involves selecting a similar problem, which can be retrieved from memory, or located as a worked example from notes or a textbook. Alternatively, if a number of similar problems have been solved before, the student may have constructed an appropriate problem-solving schema (Marshall, 1995a), in which case the student may refer to his or schema, rather than a particular problem. Unlike retrieving a particular problem or referring to a worked example, retrieving a schema may not be a deliberate, wholly conscious process. This phase is designed to support transfer and requires conceptual understanding, strategic competence, and adaptive reasoning. Holyoak and Thagard (1995) argued that transfer is likely to occur when a student successfully makes an analogy between the *source analog* and the *target analog*. According to Holyoak and Thagard, there are four steps in the use of analogy: (a) *selection*, (b) *mapping*, (c) *evaluation*, and (d) *learning*. Interpreted in this framework, Polya's second phase involves the first three steps of analogy use. During the selection stage, the solver identifies a candidate source analog (in this case, a related math problem). During the mapping stage, the solver determines the nature of the relationships between the source analog and the target analog (in this case, the math problem at hand). During the evaluation stage, the solver assesses whether or not the analogy is suitable. Even if a suitable source analog has been selected, the solution from the source analog may need to be adapted before it can be applied to the target analog.

Since Polya's third phase involves execution, it draws heavily on both strategic competence and procedural fluency. It also requires justification and verification, however, so adaptive reasoning is also used. It could be that if the student gets stuck during this phase, he or she has to reconsider which familiar problem has been selected, or, has to modify how the solution has been adapted to the target problem. The third phase involves the evaluation step from Holyoak and Thagard's (1995) framework.

In Polya's final phase, the student reflects on the solution. This includes verifying the result and consolidating the information so that it may be generalized to new problems, both of which require conceptual understanding and adaptive reasoning. This constitutes learning, the fourth step in the use of analogy. Polya's final phase can also be interpreted with respect to schema theory. If all has gone according to plan, and the problem has been solved correctly, the student will hopefully use it to enhance an existing schema or to develop a new

schema. Since powerful schemas are the hallmark of expertise, Polya's fourth phase is critical to the development of expertise in mathematics.

There is no doubt that diligent application of Polya's four phases of mathematical problem solving requires productive disposition, the fifth strand of proficiency from *Adding It Up* (Kilpatrick et al., 2001). Fortunately, the process is iterative and provides opportunities for self-correction: If a student is not careful in formulating the question in the first phase, he or she may still recover during the third phase, where each of the solution steps must be justified and verified. If a contradiction is encountered, the student will hopefully revisit the solution method, or even reread the problem to make sure it has been understood.

Following this fairly lengthy discussion of the components of mathematical problem solving, it is worth asking the question whether these are just useful constructs for describing how experts usually solve mathematics problems or whether they are actually distinct skills that, when taught to students, result in improved performance and increased transfer. For example, Kyllonen and Christal (1990) found that quantitative reasoning and working memory are highly correlated. Tirre and Pena (1993) designed a study to explore how well quantitative reasoning was explained by word problem solution skills and general cognitive abilities. They defined structural models for word problem solution skills and for general cognitive abilities. The word problem solution model had word problem identification (PI), word problem decomposition and sequencing (DS), and word problem translation (PT) components. The general cognitive abilities model had working memory (WM), verbal comprehension (VC), and reasoning (R) components. There were two quantitative reasoning variables: arithmetic reasoning (AR) and math knowledge (MK).

Tirre and Pena (1993) explored how well quantitative reasoning was fit by each of the structural models, how the structural models related to each other, and which components uniquely accounted for quantitative reasoning performance. They found that the word problem solving components did account for quantitative reasoning performance in addition to what was explained by general cognitive abilities. PT, PI, and DS all had a role beyond general cognitive abilities in predicting AR. Also, PT was not related to any of the general cognitive abilities. Of the word problem solving components, only PI had a role beyond general cognitive abilities in predicting MK. Their results suggest that quantitative reasoning

skills tap something more than general cognitive abilities, even though general cognitive abilities are related to quantitative reasoning performance.

### *Modeling and Representation*

Sigel (1999) gave the following definition for representation: "In sum, a representation refers to instances that are equivalent in meaning and in class membership, but different in mode of expression" (p. 4). The word *representation* may be used to refer to either an internal or external construct. It may also be used to describe the act of developing such a construct. All of these interpretations are included in the definition given in the *Principles and Standards* (NCTM, 2000).

Gitomer and Steinberg (1999) argued that domain considerations should inform what kinds of representations are appropriate to use in the context of a particular assessment. Alternate representations have been present in mathematics assessments for a long time, but they have not always had such explicit focus. The heightened focus is probably in response to the research that suggests students have difficulty moving between representations. By definition, a student understands a concept more completely when he or she can recognize or produce an equivalent representation. A mathematics assessment that is designed in accordance with modern standards should encourage a student to develop flexibility with recognizing and using alternate representations while assessing his or her proficiency at doing so.

In their 1992 chapter, Lesh and Lamon noted that traditional content by process matrices of standards had been criticized, and that it would be important to develop standards consistent with cognitive objectives (Greeno's term, as cited by Lesh & Lamon). Their proposed solution for responding to this criticism was to define cognitive objectives that include working with models as an explicit goal, and they noted that the 1989 NCTM process standards corresponded very well to cognitive objectives related to working with *models*. According to Lesh and Lamon (p. 26):

> Mathematical models are complete functioning systems, which consist of: (i) *elements* (for example, quantities, ratios of quantities, shapes, coordinates), (ii) *relationships* among elements within the system, (iii) *operations* or *transformations*

on elements in the system, and (iv) *patterns* that govern the behavior of the relations, operations, and transformations.

The *Principles and Standards* (NCTM, 2000) highlight the importance of models (the word model or a variant of it appears in the document 302 times, usually in the sense described by Lesh & Lamon). Also, the terms *model* and *representation* are highly related (and are sometimes used interchangeably), and representation is a distinct process standard. The term *representation* is more general, and as noted by Sigel (1999), it remains to be discovered whether it is meaningful to consider representational competence in general or whether it is strictly domain specific.

Lesh and Lamon (1992) outlined six cognitive objectives related to working with models. We will consider the first three here, as follows:[2]

- Students should use models to interpret real-life situations
- Students should think about underlying models
- Students should explore similarities and differences among alternative representation systems associated with a given model (p. 32)

The first objective pertains to model construction, which is subsumed by the representation standard. This objective is concerned with describing a real-life situation in mathematical terms (the model), so that it may be operated on. The second objective is largely concerned with evaluating the model (examining assumptions, assessing fit, and so forth). The third objective emphasizes flexibility with alternate representations, which is strongly emphasized in the standards, particularly where algebra is concerned. Consider the following quote from the *Principles and Standards* in the context of discussing linear equations:

> In the middle grades, students often begin with tables of numerical data to examine a pattern underlying a linear function, but they should also learn to represent those data in the form of a graph or equation when they wish to characterize the generalized linear relationship. Students should also become flexible in recognizing equivalent forms of linear equations and expressions. This flexibility can emerge as students gain experience with multiple ways of representing a contextualized problem. (NCTM, 2000, p. 282)

The earlier quote about representing functions (in the algebra section), makes a similar point about the importance of representation. Flexibility in the use of representations may enhance learning as well. Goldin (1998, p. 158) argued that learning mathematical concepts via multiple representations may have long-term memory benefits, both from a retention and a savings perspective, because an idea that has been encoded in multiple formats is easier to recall or to relearn.

The work described earlier by Lehrer and Schauble (2000) focused on modeling in mathematics in science. The examples they described illustrate how students improve their models over several iterations—through class discussion, the models (Lehrer and Schauble refer to the physical instantiations of the models as *inscriptions*) are reformulated and revised. An important characteristic of this work is that the students had multiple opportunities to discuss and revise their inscriptions, and to compare alternative inscriptions. This is what mathematicians and scientists do in practice. It is often argued that assessment tasks should reflect realistic content and that this is necessary, but probably not sufficient. There should be forms of assessment that incorporate realistic forms of practice (or that are seamlessly integrated with it).

### *Argument and Justification*

Mathematical argument and justification has close connections with several of the NCTM process standards and the proficiency strands from Adding it Up. It is probably most closely connected with the reasoning and proof process strand from the NCTM Principles and Standards. Recent conceptualizations of mathematical reasoning extend beyond formal proof (NCTM, 2000; Kilpatrick et al., 2001) and include the following:

- Providing examples which satisfy a statement
- Providing counterexamples
- Proposing conjectures
- Deductive argument (including forms of proof)
- Evaluating reasonableness of results, plausibility
    - Checking an answer via substitution
    - Estimation and approximation

    - Real-world knowledge
- Justification/explanation of a solution procedure

These component skills constitute forms of mathematical argument. It should be noted that the term argument has a different interpretation in mathematics than in other domains, since mathematical argument has its foundation in logic, and "opinions" should be interpreted as conjectures to be evaluated. It has been observed that an empirical approach to mathematical argument has pedagogical appeal (Schoenfeld, 1994). Developing conjectures based on examples is an important skill to develop. Successful use of an empirical approach involves understanding how examples may be used, since demonstrating through the use of examples is not always sufficient (e.g., Epp, 2003; Weber, 2003). Attempting to "prove" the truth of a statement through the use of an insufficient set of supporting examples is a common error, as evident in the earlier discussion of the Lee and Wheeler (1989) study.

Michener (1978) developed a classification scheme for examples with respect to their role in teaching and learning. Among the types of examples included were "start-up" examples (to facilitate understanding), and counterexamples, in addition to more standard kinds of examples. Understanding the different types of examples may help students to consider counterexamples as well as supporting examples, and to construct more sound mathematical arguments. Asking students to justify a correct response is a common task, but having them also explain why incorrect answers are incorrect may be even better (Stigler & Perry, as cited in Siegler, 2003).

Mathematical argument and justification are also closely tied to the NCTM communication and connections process strands. Mathematical argument is the central part of mathematical communication—often, what one is communicating in mathematics is some form of argument, at least as argument has been broadly defined here. A student who communicates an effective argument not only presents it accurately, but explains it clearly, with attention to the target audience. In a mathematical argument, accuracy and clarity are often inseparable, and it may be impossible to interpret the accuracy of an argument that is not clearly explained. This is not to suggest that mathematical arguments must be lengthy in order to be clear—some of the best arguments are also the most concise.

There is also a close connection between representation and mathematical argument. In order to construct a sound mathematical argument, a student must identify an effective

representation. As suggested from the Lee and Wheeler (1989) study, sometimes the most effective representation to support or refute a statement is different from the representation in which the statement is given. In making a mathematical argument, students may also draw on a number of different representations, particularly in arguments that require multiple steps.

***Summary***

In this section, key processes central to mathematics learning were described, including problem solving, representation, and mathematical argument and justification. Although they have been discussed in separate sections, these processes are highly interrelated. For example, students must identify appropriate representations in order to solve a problem or develop an argument. Where problem solving is concerned, a goal for assessment should be to help students develop a systematic approach for identifying and organizing relevant information, and to encourage the development of gradually more general methods of solution. Mathematics assessments should also encourage students to identify appropriate representations and to use them flexibly in developing clear and accurate mathematical arguments.

## First Draft of Competency Models for Middle-School Mathematics

### *A Model of Competency With Respect to Mathematics Content*

The earlier portion of this document characterized core content and key processes in mathematical competency, with the goal of informing development for cognitively based assessments for middle-school mathematics. At this point, we are ready to consider how to draft a model for middle-school mathematics content and process competency; we consider content first. The K-12 mathematics curriculum spans a great deal of content, and it is a concern to many educators that the coverage is too broad and lacks sufficient depth. In an attempt to respond to this, the NCTM released the *Curriculum Focal Points for Prekindergarten Through Grade 8 Mathematics* (2006). This document highlights core mathematics content by grade. In each grade, there are three central content areas, or focal points. The document also specifies two kinds of connections to the focal points within each grade: (a) connections from focal points in other grades, and (b) connections from less central content, within the same grade. Thus the curriculum focal points assume a model of student acquisition.

The first draft of the competency model for content is based on the NCTM curriculum focal points. Figure 1 is a graphical interpretation of the focal points (and the connections to them) for grades 6-8. Text descriptions of the focal points and connections to them are given on pages 18-20 of the focal points document. The graphic in Figure 1 does not provide the same level of detail that is provided in the text descriptions; its purpose is to make the focal points and connections to them visually apparent at a high level.

The graphic in Figure 1 is divided into three columns, with grade 6 on the left, grade 7 in the middle, and grade 8 on the right. The focal points are shown in the boxes with solid red borders. The text in these boxes is from the description headings given in the NCTM *Curriculum Focal Points* (2006), though headings are shortened to save space. Connections to the focal points are represented by the boxes with dashed borders; the original text descriptions are quite lengthy, and there are no description headings, so these are paraphrased from the original document.

Each focal point and connection is identified with at least one content strand from the NCTM standards (NCTM, 2000); this is shown in each box and abbreviated as follows: numbers and operations (N&O), algebra (A), measurement & geometry (M&G), and data analysis and probability (DA&P). Measurement and geometry were consolidated, because for grades 6-8, focal points or connections to them identified by M&G were related to both content strands, with the exception of one case. The boxes are color coded in accordance with the corresponding content strands. Cyan, magenta, and yellow are used for algebra, numbers and operations, and measurement and geometry, respectively. White is used for data analysis and probability. Focal points or connections that draw on multiple strands are filled with mixtures of these colors.

The arrows represent relationships between focal points, and between connections and focal points, as inferred from the document. Note that the arrows always lead to a focal point (they only lead away from focal points when leading to other focal points). These arrows, of course, do not represent all connections among content areas, but are designed to highlight how connections are related to focal points, and how focal points are related to each other.

The two types of connections discussed earlier are evident from Figure 1. First, consider connections across content but within grades. For example, in grade 6, *Finding areas and volumes* (measurement and geometry) is not a focal point. But it is a context in

which students *Write, interpret, and use mathematical expressions and equations* (algebra), which is a focal point. In general, within-grade connections to a focal point represent alternate contexts in which the focal point may be addressed. Next, consider the across-grade connections, which tend to be cumulative. For example, *Understanding and application of proportionality* . . . is preceded by *Connecting ratio and rate.* Since a proportion is an equivalence between ratios, students must understand the concept of ratio before understanding the concept of proportion.

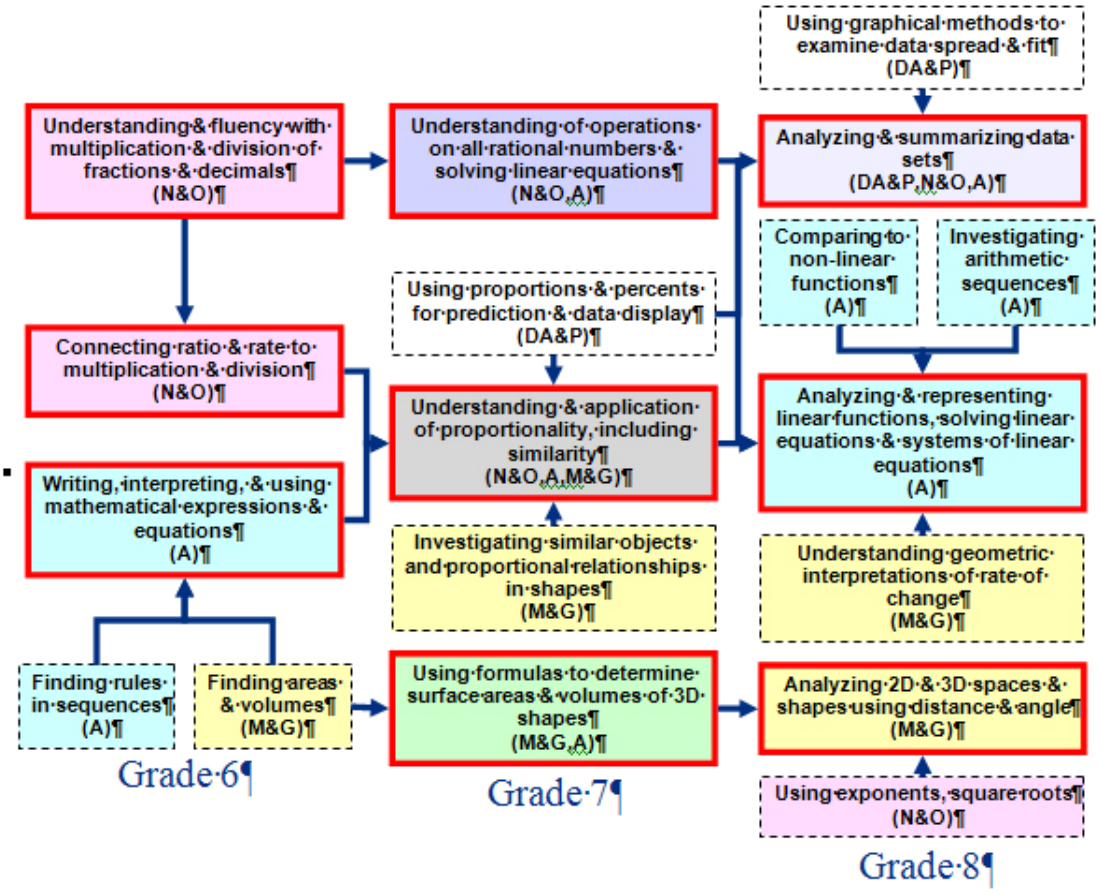


***Figure 1*. Graphical interpretation of the Curriculum Focal Points for Prekindergarten Through Grade 8 Mathematics (NCTM, 2006, pp. 18-20).**

A couple of other features of Figure 1 are worth noting. First, of the nine focal points represented in grades 6-8, six involve algebra; five involve numbers and operations; three involve measurement and geometry; and one involves data analysis and probability. This

would tend to support the idea of focusing on algebra and numbers and operations, and the connections between them, in the middle grades. As mentioned earlier, however, these areas should not be given emphasis to the exclusion of other important content. Second, note that proportional reasoning is a "strongly connected" focal point that corresponds to several content strands. It is linked both within and across grades. As discussed earlier, it is also an area that has traditionally posed difficulty for students.

### *A Model of Competency With Respect to Mathematics Process*

Three processes were discussed in the section on mathematics competency: *problem solving*, *representation*, and *mathematical argument and justification*. While it is widely used, the term *problem solving* is extremely broad and has many different interpretations. Sometimes it refers to the process of solving word problems. Sometimes it is used much more generally and refers to almost any aspect of dealing with a mathematical problem, including representing the situation, planning the steps, executing the procedures, verifying the result, and so on. For this reason, it may not be very meaningful to develop a competency model for problem solving, at least not without breaking it down.

Figure 2 shows a draft model for process competency in mathematical argument and justification (top) and representation (below). Since the development and use of these two processes is concomitant, it seemed reasonable to put them on a common timeline. First, consider the development of mathematical argument and justification. It is assumed that students first learn to provide examples that support an argument. This is not to suggest that positive examples are always easier to find, just that they are easier to reason about. Next, students may learn to identify falsifying cases and counterexamples; this would be followed by informal methods of direct proof, and finally by a variety of formal proof methods. This proposed sequence of stages is very loose, and it is expected that there would be significant interactions with task type. For example, some false statements may have counterexamples that are difficult to identify, and some true statements may be very straightforward to verify directly. Similarly, although students may be more comfortable using simple language to make mathematical observations early on, developing an extended verbal argument is more demanding than producing a simple chart or table. The sequence proposed in this model reflects conjecture and needs to be investigated in further research.

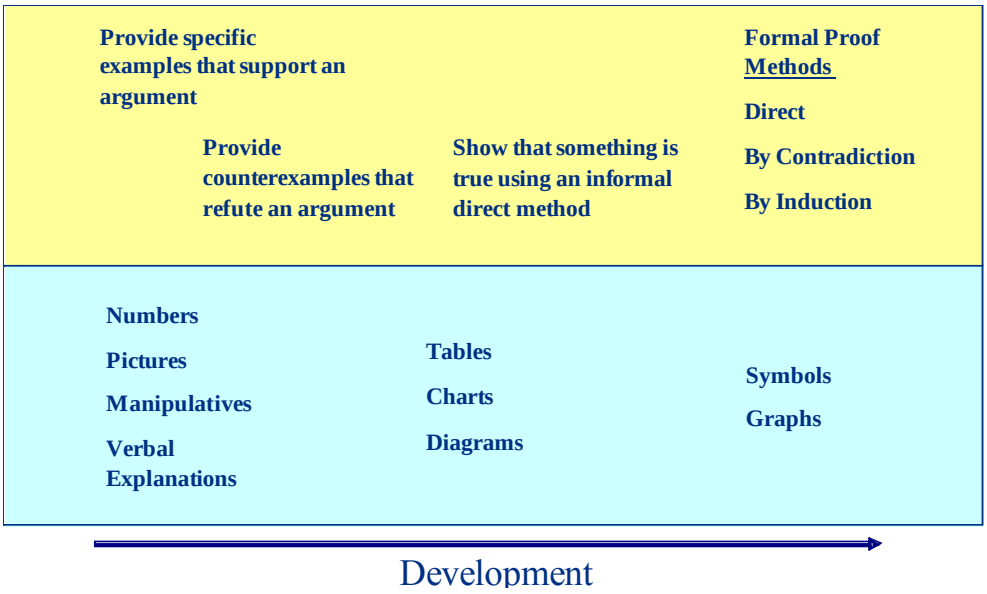


***Figure 2.*** **Draft competency model of mathematics process.**

Next consider the development of supporting representations, shown in the lower portion of Figure 2. In a review of the influence of cognitive psychology on mathematics education, English and Halford (1995) summarized some of the work of Jerome Bruner and Zoltan Dienes. Bruner and Dienes suggested that children's skill with using representations develops from the more concrete to the more abstract. As cited in English and Halford, Bruner suggested that representational skills develop in three stages: *enactive*, *iconic*, and *symbolic*, and Dienes proposed that children begin by using concrete materials and progress to more abstract representations such as pictures and eventually graphs and symbols. This progression, from the more concrete to the more abstract, is reflected in Figure 2.

Note that there is an implied difference in the time span between Figures 1 and 2. Figure 1 focuses on the middle grades, whereas Figure 2 spans development from early childhood through high school. This partly reflects that content learning is largely curriculum driven, and what content is covered at each grade level is more highly specified than which processes occur at each grade level. The scope of Figure 2 could be restricted to cover only middle school, but since process development is necessarily loose, the model would probably be less meaningful.

The models discussed in this section constitute an initial attempt to represent the Curriculum Focal Points and findings from the research literature in graphical form. They are

discussed here primarily to describe the rationale that led to the design of the current model of mathematical competency that is used for CBAL mathematics task development. This model is still evolving, but is presented and discussed at length in Graf, Harris, Marquez, Fife, and Redman (2009).

## Describing and Quantifying Evidence of Mathematical Proficiency

As mentioned at the outset, it is anticipated that both an accountability assessment system and a formative assessment system might share a common model of mathematical competency. The evidence models may differ substantially between these two types of assessment systems, however, since they will be used in different contexts and will have different applications.

The discussion of evidence is organized into four main sections: developmental progressions, strategies, bugs and misconceptions, and the role of the situative perspective. Research in each of these areas that is relevant to the question of evidence is summarized. Although the focus of each of the four sections is different, they interact to a high degree. As their mathematical thinking develops, students use different strategies. The use of particular strategies can lead to different kinds of bugs, and misconceptions can influence the choice of strategy. Similarly, many strategies and misconceptions are situation-specific.

### *Developmental Progressions*

Although they span different time frames, both the content and process competency models imply a developmental progression. Particularly where formative assessment is concerned, it is important to consider prerequisite competencies for the material under study, as well as what the student should be able to do at the next stage. The Sfard (1991) work discussed earlier posits a general model for the development of mathematical ideas. Sfard argued that operational (process-based) conceptions generally precede structural (object-based) conceptions, both in the historical development of mathematical ideas and within an individual learner. Sfard's view is expressed in the following quote:

> Of the two kinds of mathematical definitions, the structural descriptions seem to be more abstract. Indeed, in order to speak about mathematical *objects*, we must be able to deal with *products* of some processes without bothering about the processes themselves. In the case of functions and sets (in their modern sense) we are even

> compelled to ignore the very question of their constructivity. It seems, therefore, that the structural approach should be regarded as the more advanced stage of concept development. In other words, we have good reasons to expect that *in the process of concept formation*, *operational conceptions would precede the structural*. Different kinds of evidence will be brought in this article to show that this statement is basically true whether historical development or individual learning is concerned. (p. 10)

One of the examples discussed by Sfard (1991) concerned the development of the function concept. According to Sfard, early conceptions of function were closely tied to algebraic representations that emphasized the role of variables. A later definition from Euler (as cited in Sfard, p. 15) did not mention variables, but emphasized the notion of dependency—that is, that one quantity changes with respect to another. Sfard considered these early definitions operational because they were expressed in terms of processes applied to variables or quantities that resulted in other variables or quantities. Eventually these definitions were subsumed by Bourbaki's definition, which Sfard considered structural, because it describes the function concept with respect to a set of ordered pairs rather than with respect to an operational process.

Sfard (1991) suggested that there are three stages in the development of a mathematical concept: (a) *interiorization*, (b) *condensation*, and (c) *reification*. During interiorization, the learner becomes familiar with operational processes and computations. In condensation, the learner begins to consolidate operational steps, and the focus shifts from the details of a procedure to its result. Finally, during reification, the learner begins to use the concept as a structural object that can itself be applied to the interiorization of new concepts. The Sfard framework may be applied to the developmental progression for any mathematical concept, but, of course, research specific to the concept must inform its developmental progression as well.

Progressions of mathematical understanding for infants and very young children are relatively well understood (see Siegler, 2003, for a review). This may be partly because very early mathematical concepts develop prior to any formal instruction. The development of student strategies and schemas for solving arithmetic word problems has also been extensively investigated (Briars & Larkin, 1984; Carpenter & Moser, 1984; Morales, Shute,

& Pellegrino, 1985; Riley, Greeno, & Heller, 1983). As students learn more advanced material, their experiences with mathematics learning and instruction vary to a greater extent. Also, as the material becomes more complex, there are more solution strategies that may be brought to bear on any particular problem, so measures of performance become increasingly complex as well. By necessity, studies of how mathematical skill develops at the later stages are more specific to a particular content area or even to a particular task type.

Noelting (1980) investigated the development of students' understanding of proportional reasoning comparison problems. This work is an excellent example of how to coordinate measurement and task design, so that the results may be interpreted with respect to a developmental model. The child was shown two pitchers. Then, in each pitcher, the experimenter mixed some number of glasses of water and some number of glasses of orange juice. The task was for the child to decide which pitcher would taste more strongly of orange juice. Noelting used the notation ($a$,$b$) to represent the mixture in a pitcher, where $a$ is the number of glasses of orange juice, and $b$ is the number of glasses of water. A number of different ratio comparison tasks were systematically developed. Tasks as simple as ($a$,0) versus (0,$b$) (in one mixture only orange juice is present, in the other mixture, only water is present), and as complicated as (3,5) versus (5,8) were included, as were tasks intermediate in difficulty. For example, (3,4) versus (2,1) was such an intermediate pair.

Noelting (1980) organized the tasks into seven stages according to difficulty, and characterized the tasks at each stage with respect to their cognitive demands. For example, to solve an ($a$,0) versus (0,$b$) task, only "identification of elements" (Table 5, p.231) is required. The (3,4) versus (2,1) task is characterized by "an inverse relation between terms in the ordered pairs" (Table 5, p. 231). Note that neither of these tasks requires a formal approach such as finding common denominators, but that the (3,5) versus (5,8) task is not easily solved using such a weak method. Noelting observed that students who could solve tasks at higher stages were generally older, and he interpreted this observation with respect to a Piagetian framework. He performed a confirmatory factor analysis and extracted six factors instead of seven, but found that the correspondence between the proposed seven stages and the extracted factors was quite good.

Modern views on developmental progressions recognize that skills and competencies are not necessarily associated with a particular age band (at least past a certain age), and I

will take that perspective here. There have been several attempts to identify "progress maps" or "prerequisite maps" in accordance with developmental progressions in mathematics, a few of which will be discussed here.

Project 2061 (American Association for the Advancement of Science [AAAS], 2001) has developed the *Atlas of Science Literacy*, which includes a set of "conceptual strand maps" for different topics in science and mathematics. There are maps for both process strands and content strands, and each map shows the relationships among the proposed skills and concepts in each map. Volume 1 includes maps on mathematical processes, mathematical models, graphic representation, symbolic representation, ratios and proportionality, and describing change. A sample map for ratios and proportionality can be accessed from http://www.project2061.org/publications/atlas/sample/9_3_RP.htm.

Volume 2 of the *Atlas of Science Literacy* and draft maps are available from http://www.project2061.org/publications/atlas/sample/toc2.htm. According to the Web site, the new volume includes maps on mathematical applications, shapes, and reasoning. The *Atlas of Science Literacy* maps have a cognitive research basis, but these appear to have been developed primarily for pedagogical purposes rather than for formal assessment.

In an effort to realign achievement tests to a new elementary school curriculum adopted in the Philippines, the Center for Educational Measurement developed a set of progress maps for mathematics (Angeles, Sampang, & Moseros, 2006). The progress maps cover grades 1-6, and the manuscript provides the progress map developed for fractions as an example. A distinguishing feature of this progress map is that interrelationships among content area skills are represented both vertically within each grade level and horizontally across grade levels—it is similar to the NCTM curriculum focal points in this respect, although the skills are described at a finer grain size. The progress map was developed by a mathematics curriculum expert, in accordance with test specifications, and was reviewed by other mathematics experts in an initial validation effort. The paper also includes examples of lessons that were developed to address each skill represented in the progress map. A longitudinal study to validate the sequence of skills in the progress map is planned.

Falmagne, Cosyn, Doignon, and Thiéry (2006) described the use of knowledge spaces theory as it is applied to the ALEKS adaptive assessment system for mathematics. ALEKS relies on large precedent maps (for example, there is one for algebra) that specify prerequisite

relationships among different problem types. Any minimum set of precedent problem types that imply mastery of another problem type is referred to as a *knowledge space*. Knowledge spaces are built by asking experts to determine whether a student could solve a particular problem, given that he or she was unable to solve some other combination of problems. The system adapts by giving the student problem types from the "outer fringe" (immediately beyond the likely knowledge state) when the student is doing well, and retreating to problem types from the "inner fringe" (just below the likely knowledge state) when the student is having difficulty. Once the likelihood distribution for a student's knowledge state reaches some minimum entropy, the system estimates the final knowledge state, administers a final problem, and stops. ALEKS' knowledge structures are validated by comparison against actual student response data. According to the paper, ALEKS is highly accurate—the correlation between responses as predicted by the final knowledge state and the observed responses is between .7 and .8.

It would be interesting to explore task difficulty factors from a developmental perspective (I. Bejar, personal communication, April 21, 2009). In other words, it is likely that different task features influence difficulty in different ways at different stages of development. One way we can begin to build a developmental framework is to investigate which task features influence difficulty across grade levels.

### *Strategies*

Implicit in the earlier discussion about alternate representations and the development of mathematical problem solving and reasoning is the assumption that students may use alternate strategies in solving mathematics problems. This poses a challenge for assessment development, because different solution strategies may reveal very different kinds of evidence. Although it is not based on middle-school content, an example from Schoenfeld (1987) illustrates how this can happen. He presented students with the task of finding an integral. It was intended as a warm-up task, because the integral can be found using a simple substitution method (Schoenfeld estimated that applying this strategy should take students about 2 minutes). What he observed, though, was that a number of students solved the same task using either a partial-fractions method or a trigonometric substitution. He noted that while both of these methods require greater knowledge of mathematics, they take much more time to complete. His point was that metacognitive monitoring and strategy selection is

important to develop in mathematics problem solving. While it is valuable for students to try alternative solution methods, ideally they should also develop skill with recognizing efficient solutions.

From an assessment perspective, strategy by task interactions pose challenges for both task design and scoring. One way to assess use of a particular strategy is to constrain the task in order to encourage the student to solve it in a particular way. This defeats the purpose, however, if part of the goal is to assess a student's strategic planning and whether or not he or she makes a judicious strategy selection. Another approach is to systematically design a set of tasks such that different tasks require different levels of strategic sophistication. This is the approach that Noelting (1980) used in his work on proportional reasoning comparison problems. This approach is helpful for identifying stages in strategic development. For complex tasks that can be solved using many different methods, however, more research is needed on how to help students compare alternate strategies in order to select one that is efficient.

Strategy by task interactions may be difficult to accommodate on the scoring side as well. For example, a student might earn points for selecting an efficient strategy or for choosing a strategy that reflects a high level of mathematical understanding. It may not be appropriate to combine these points into a composite score—but they can't be considered entirely separately either. For example, it may be useful to know that a student's strategies tend to reflect a high level of mathematical knowledge, even if the student tends to apply inefficient strategies. But the student who always selects the most efficient strategy (that may also require a lower level of mathematical knowledge) may have a high level of mathematical knowledge that is not in evidence. One possible way to resolve this is to design some tasks that primarily assess strategic efficiency and other tasks that primarily assess mathematical knowledge.

Mayer, Larkin, and Kadane (1984) described an experiment that explored the influence of representation, or task format, on strategy. They developed isomorphic pairs of single-variable algebra problems: one member of the pair in equation format and the other member of the pair in word problem format. In earlier work, they observed that students used two strategies for solving such problems, which they referred to as the *reduce strategy* and the *isolate strategy*. Students who used the reduce strategy initially focused on simplifying

expressions on either side of the equation; students who used the isolate strategy initially focused on moving the variables to one side of the equation and the numbers to the other. The isolate strategy is more complicated, because it is not always possible to immediately isolate a variable without first doing some simplification—resulting in what Mayer et al. referred to as *goal stacking*.

Each equation (or word problem) was decomposed into problem states (where each state corresponded to a particular stage on a possible route to solution). Students solved the problems working forward from each of the possible states. Each strategy type implied different sequences of actions, which in turn predicted different patterns of response times across problem states. Based on the fit between the strategy predictors and the observed response times, Mayer et al. (1984) concluded that students appeared to use the reduce strategy to solve word problems and a more complicated goal-stacking strategy to solve equation problems.

Proportional reasoning is another area where researchers have identified many alternative solution strategies. In solving missing value proportional reasoning problems, Vergnaud (1983) observed that students prefer to apply *scalar* strategies rather than *functional* strategies. Use of a scalar strategy involved noticing the transformation within one *measure space* and applying it to the other measure space. For example, consider this problem: "If 3 yards of ribbon cost 15 cents, how many cents do 9 yards of ribbon cost?" Here, yards of ribbon define one measure space and cents define the other. Students generally prefer to note that 9 yards equals 3 times 3 yards, so the cost must be 3 times 15 cents, or 45 cents. This is a scalar strategy. But the problem may also be readily solved by working across measure spaces—since 15 is 5 times 3, the cost for 9 yards must be 5 times 9, or 45 cents.

Kaput and West (1994) also identified a number of different strategies for solving proportional reasoning problems, including *build-up processes* and the *unit factor approach*. Build-up processes include strategies such as *coordinated double-skip counting* (Kaput & West). For example, if 3 yards cost 15 cents, then 6 yards cost 30 cents, and 9 yards cost 45 cents. In this example, the unit factor approach involves finding the price per unit (5 cents/yard) and multiplying by the number of units (9 yards) to yield the result, 45 cents.

Across areas in mathematics, multiple solution methods are possible, and different strategies reveal different evidence with respect to both mathematical content and process

knowledge. To complicate matters further, the format in which a task is presented can influence which strategies are preferred. Also, strategies are not necessarily pure instantiations of a particular approach; they may consist of hybrid solutions and false starts. Hall, Kibler, Wenger, and Truxaw (1989) developed a classification scheme for strategic episodes in the solution of common algebra word problems. The strategic episodes identified in this scheme were not necessarily mutually exclusive; a student might draw on several of them during the course of solving a problem. Some of the strategic episodes were related to understanding and representing the information given in the problem, others were related to alternative correct approaches, and still others were related to different types of errors. Such a classification scheme is useful because lengthy solutions can be coded and compared with respect to their constituent strategic episodes.

It is probably clear from the preceding discussion that strategy by task interactions are important to consider for the purpose of interpreting evidence and must be attended to in the design of sound assessments. It is probably also clear that strategy analysis in mathematics is both labor-intensive and content-specific. It is probably not feasible to identify each different strategy for each type of task for each topic in middle-school mathematics. Nor would information at this level of detail necessarily be useful to report to stakeholders. It may be possible to consolidate information about strategies at a high level. The question is whether it is possible to do this accurately without *first* conducting the fine-grained, detailed analysis.

Stevens and Thadani (2006) described an approach for categorizing different strategies students used as they solved problems from the Hazmat problem set. Hazmat consists of problems in which there has been a toxic spill, and the student's task is to identify what substance has been spilled by selecting a variety of tests. There are multiple routes to solution. Not all are correct, and those that are correct are not all efficient—for example, students may conduct more tests than necessary. Stevens and Thadani used an artificial neural network to identify strategies based on the relative frequencies with which different tests were selected. Similar strategies were clustered into a small number of states (in this case, five states). They then used hidden Markov models to estimate transition probabilities among the five states. They noted that while this approach is very useful for research, the models are different for each problem set, and teachers need strategy information summarized at a higher level. So performances were categorized with respect to both

strategic efficiency and outcome, as follows: high efficiency, low outcomes; low efficiency, low outcomes; high efficiency, high outcomes; and low efficiency, high outcomes. This sort of approach could provide teachers with high level information about students' strategies and how they might be modified.

### *Bugs and Misconceptions*

Bejar (1984) distinguished between diagnostic assessment that involves deficit measurement and diagnostic assessment that involves error analysis. This section is concerned with the latter, where errors are defined as bugs and misconceptions. First, to distinguish the two: a bug is a procedural lapse and may be just a slip but, as discussed earlier, it may also indicate conceptual misunderstanding. Bugs (and their diagnostic limitations) have been thoroughly explored in some mathematical areas, in particular, basic arithmetic (e.g., Brown & Burton, 1978; Van Lehn, 1983) and solving linear equations (e.g., Payne & Squibb, 1990; Sleeman, 1984). Facet-based instruction, or FBI (Hunt & Minstrell, 1994; Minstrell, 2001), organizes diagnostic physics items around student misconceptions, or *facets*. The items are deliberately constructed to assess the prevalence of facets in different contexts.

The research on bugs and misconceptions suggests that there are a number of bugs and misconceptions that occur with very high frequency. Some of them have been discussed earlier in this document. In proportional reasoning, for example, the "incorrect addition strategy" (Hart, 1984) reflects a misconception about the nature of proportional relationships. This misconception is pervasive and observed across contexts. The variable reversal error discussed earlier is also very common among college students (Clement, Lochhead, & Monk, 1981). This error appears not to be a mere slip, since pointing it out or even remediating it does not appear to resolve the problem, at least not in the short term (Graf, Bassok, Hunt, & Minstrell, 2004; Rosnick & Clement, 1980). The protocols from Lee and Wheeler (1989) suggest that some students do not readily perceive the connection between arithmetic and algebra and will produce explicit justifications for incorrect procedures. Other common bugs include inappropriate cancellation (Lee & Wheeler), inappropriate application of cross-multiplication (for example, when adding two fractions—see Kaput, 1999), and difficulty in operating on expressions with parentheses (Lee & Wheeler; Payne & Squibb, 1990; Sleeman,

1984). While many of the errors may be due to slips, many are likely due to an incomplete understanding of the system.

Even though there are certain bugs or misconceptions that occur with reasonable frequency, a particular student does not necessarily adhere strongly to a particular misconception or naïve theory, and students use different *mal rules*, or incorrect rules, even on similar items (Payne & Squibb, 1990). Madhyastha, Hunt, Kraus, and Minstrell (2006) investigated the "coherence" of student response patterns on items dealing with forces and motion. *Coherence* refers to how consistently students respond in accordance with the same facet, or misconception, across similar items. They made several interesting observations. First, there were a number of prevalent misconceptions. Nevertheless, on a pre-assessment, only 35% of students showed coherent patterns of responses, suggesting the large majority were not operating in accordance with a particular theory. Following instruction, on a post-assessment, 44% of students showed coherent (though not necessarily correct) patterns of responding. Coherence improved with instruction, but the majority was still not responding in accordance with a coherent pattern. Madhyastha et al. also observed that coherent patterns of responding appeared to be correlated with math ability. They interpreted this to mean that there may be a relationship between coherence and sophistication in reasoning about the material. Similarly, in a study investigating students' patterns of responding to algebra items, Payne and Squibb (1990) observed that while students generally respond with inconsistent patterns of algebra mal rules, it is easier to diagnose students with greater levels of algebra skill.

When misconceptions are diagnosed for a particular student, there is an implicit assumption that the student is responding in accordance with an idea. But these results suggest that students do not necessarily respond consistently, perhaps particularly when they are not comfortable with the material. These results have implications for how evidence is accumulated. For misconceptions that are common, it is probably worthwhile to target the misconception at the class level, since chances are good that some proportion of students will apply it at one time or another. But it also suggests that we should not attempt to diagnose individuals on the basis of their responses to only one or two questions. If a student responds consistently with a popular misconception, then it is appropriate to diagnose at the individual level. A generic diagnosis could still be provided at the individual level for inconsistent

patterns of responding—perhaps the most appropriate course of action would be to refer the student to the teacher for additional help.

Recall the example presented earlier from Harel (as cited in Kaput, 1999), where the problem solver responded to superficial characteristics of the task while trying to solve an inequality. To characterize all of the different possible superficial configurations in order to build an exhaustive model of bugs and misconceptions would be a formidable task. By focusing on the improvement of students' foundational skills, it is possible that some of these bugs and misconceptions will disappear, as students develop a deeper understanding of the system in which they are operating.

In a discussion about student misconceptions, Wiliam (2007) pointed out that while incorrect responses can reveal important evidence, what is most important is that the correct response is interpretable. If it is possible to obtain the correct answer without understanding, then the item does not provide sufficient evidence for student understanding. In assessment contexts with high-stakes outcomes, items that may be answered correctly through use of construct-irrelevant strategies often manifest as poorly discriminating during pretesting and are subsequently eliminated. Note though that even an item that discriminates well does not necessarily assess deep understanding—that is accomplished through careful attention to task design, which will be discussed later in the document. The issue to note here is that an argument can be made that while a lot of research has focused on the nature of bugs and misconceptions, less attention has been directed to the interpretability of correct responses, and this may actually be the most important place to focus our attention.

In sum, it is recommended that we track and remediate the most common misconceptions, though not at the individual level, unless there is clear evidence for a consistent pattern of responding that we can identify. Bugs and misconceptions can suggest weaknesses in foundational skills. As noted by Wiliam (2007), however, attention to the interpretability of correct responses is most important, so it is not advisable to focus on bugs and misconceptions to the exclusion of this consideration. A potential focus for future research is to determine whether it is more effective to remediate by developing foundational skills or by explicitly addressing bugs and misconceptions.

### *The Role of the Situative Perspective in Mathematics Assessment*

During the 1990s, there was a spirited debate between researchers who argued that learning theory should be approached from a cognitive perspective (e.g., Anderson, Reder, & Simon, 1996; Anderson, Reder, & Simon, 1997) versus researchers who argued that a situative perspective was also important (e.g., Greeno, 1997). The cognitive perspective is focused on how a learner infers meaning, represents information, and solves problems, while the situative perspective is focused on how learning practices are cultivated in the context of the surrounding environment. In quantitative domains, research from the cognitive perspective has focused on identifying solution strategies of individual problem solvers, identifying components of knowledge and skill and how these components are organized into schemas, studying the role of prior knowledge and misconceptions, identifying features of tasks that influence difficulty and solution strategy, and developing interventions targeted towards improving individual learning.

Research in quantitative domains from the situative perspective has focused on how environmental experiences interact with performance, how skilled practice develops in a community setting, and the importance of task authenticity. Shute and Psotka (1996, pp. 585-586) discussed the implications of the two perspectives for the design of an intelligent tutoring system. Similarly, the two perspectives have implications for the design of mathematics assessment. The individual studies discussed to this point have focused on research from the cognitive perspective. In this section, research studies in quantitative domains from the situative perspective are described.

Since the 1990s, there has been a shift from contrasting the cognitive and situative perspectives to highlighting their common goals:

> A more complete cognitive theory will include more specific explanations of differences between learning environments, considered as effects of different contexts, and a more complete situative theory will include more specific explanations of individual students' proficiencies and understandings, considered as their participation in interactions with each other and with material and socially constructed conceptual systems. (Anderson, Greeno, Reder, & Simon, 2000, p. 12)

> Both [cognitive and situative] perspectives imply that assessment practices need to move beyond the focus on individual skills and discrete bit of knowledge that

> characterizes the earlier associative and behavioral perspectives. They must expand to encompass issues involving the organization and processing of knowledge, including participatory practices that support knowing and understanding and the embedding of knowledge in social contexts. (Pellegrino, Chudowsky, & Glaser, 2001, p. 65, brackets added)

In other words, although the cognitive and situative viewpoints emphasize different aspects of learning, they are not mutually exclusive. Where assessment development is concerned, the cognitive and situative perspectives ideally should play complementary roles.

Schliemann and Nunes (1990) examined how fishermen in Brazil solved missing-value proportion problems. The fishermen use proportional reasoning to evaluate how the price they command for caught fish compares to the price at which prepared fish is selling on the market. Only two of the fishermen in their sample had schooling through seventh grade, when procedures for solving proportions are typically taught. In Brazil, students are instructed to solve missing-value proportion problems by applying the *Rule of Three*: $\frac{a}{b} = \frac{c}{x}$, where *a*, *b*, and *c* are replaced by the known quantities in the problem, and *x* represents the unknown quantity. Students are then shown how to cross-multiply and solve to find *x*. Even though most of the fishermen had not had formal instruction on how to solve proportions, they were able to solve proportion problems effectively (including transfer problems that differ from the kinds of problems they solve in the marketplace) by using the scalar strategy discussed earlier. In a subsequent interview with students who had received formal instruction, Schliemann and Nunes again found that use of the Rule of Three algorithm was uncommon.

As part of the Concepts in Secondary Mathematics and Science (CSMS) project, student work on ratio and proportion problems was examined for frequent strategies and errors. Although the Rule of Three is a commonly taught algorithm in British schools, only 20 of 2,257 student papers used the Rule of Three approach (Hart, 1984). Schliemann and Nunes (1990) concluded that proportional reasoning does not have to be formally instructed in order to be learned, and that in fact the standard procedure taught in school may conflict with students' preferences for how to solve such problems. One of the tenets of the situative perspective is that rather than trying to replace students' intuitive (and often correct) ideas

about how to solve problems with more abstract and inscrutable procedures, instruction should build on students' existing understanding. This is not to say that formal and efficient solution procedures should not be taught—only that they should be connected to what students already understand from informal experiences.

Carraher, Carraher, and Schliemann (1985) posed mathematics problems to children between the ages of 9 and 15 who were working as street vendors in Brazil. The children were first asked mathematical problems as part of an informal interview, in the context of the children's working environment. The children were asked questions about purchases (e.g., how much will some number of units of a particular item cost). Following the interviews, children participated in a formal test that consisted of numeric computations and word problems. The formal test was customized for each child—it included only problems that used the same numbers as problems that the child had solved correctly during the informal interview.

Carraher et al. (1985) found that performance on situated problems (whether problems from the informal interviews or word problems from the formal test) was much higher than performance on the numeric computation problems. This is somewhat striking when it is considered that the interview computations were performed mentally, while during the formal test children had access to paper and pencil. Carraher et al. also noticed that the children used school-based procedures on the formal test to a greater extent than in the interviews, where they relied more on intuitive methods. They concluded that the standard approach of teaching formal mathematical methods prior to introducing contextualized problems (such as word problems) should be reconsidered.

Lave, Murtaugh, and de la Rocha (1984) compared shoppers' performance with solving arithmetic problems in the supermarket to their performance on a formal paper-and-pencil test of arithmetic. They found that performance on problem solving in the supermarket was 98% correct, while performance on the written arithmetic test was only 58% correct. But they noted that while shoppers' final solutions in the supermarket tended to be correct, their problem-solving behavior had a characteristic form: they tended to make multiple calculations and, often, the intermediate calculations contained errors. Lave et al. proffered the following possible explanation for the shoppers' supermarket problem-solving behavior: the supermarket environment and the shoppers' knowledge of it affords quite a bit of

information about what is typical and reasonable. The interaction between the shopper's experience and the information present in the environment function to establish constraints on the solution. Lave et al. gave the following description as an example of how problem solving and environment may interact:

> One shopper found an unusually high-priced package of cheese in a bin. He suspected an error. To solve the problem, he searched through the bin for a package weighing the same amount and inferred from the discrepancy between prices that one was in error. (p.77)

De la Rocha (as cited in Lave et al., 1984) gave another interesting example of using the environment to solve a problem. When asked to find three-quarters of two-thirds of a cup of cottage cheese, a Weight Watchers member measured out two-thirds of a cup of the cheese, spread out the contents and sectioned it into quarters, and removed one quarter.

The actions of the shopper and the Weight Watchers member are not so different from what children do when they use manipulatives or supporting software to reason about mathematics problems. Because they are physical objects, manipulatives may make the affordances and constraints of a system more obvious to the problem solver. The interaction between the problem solver and his or her environment is something that needs to be considered in the interpretation of evidence. It may be that students who have difficulty reasoning mathematically in formal situations can nevertheless reason quite successfully in familiar environments. As another example, Hoyles, Noss, and Pozzi (2001) found that nurses used a variety of proportional reasoning strategies to determine what doses to administer to patients. More often than not, they did not rely on the computational rule they had been taught, but they did not make any errors when administering the doses. Hoyles et al. knew from prior work, however, that nurses did not score highly on formal tests of proportional reasoning.

What this body of research suggests is that both children and adults may have situation-specific competency that is not necessarily reflected in more formal measures of assessment. It may be, however, that situation-specific skill is a precursor to more abstract and formal understanding. It may be useful to recognize evidence of situation-specific competency, because students at this stage may advance to the next level of abstraction if provided with sufficient scaffolding and support.

### *Summary*

This section focused on four interrelated areas of research—developmental progressions, strategies, bugs and misconceptions, and the role of the situative perspective—all of which pertain to the nature of evidence collected in mathematics assessment. If the goal is to develop assessments that will serve learning, they need to provide evidence that locates students with respect to a proposed developmental trajectory. Although mathematical content and process competencies are applied in combination, they have been considered separately because the development of mathematical knowledge is more prescribed and largely curriculum-driven, while the development of processes such as representation and argument are more fluid.

Many tasks afford multiple solution methods, and this implies that the same task may elicit very different evidence depending on how the student responds. This raises questions about how to deal with lengthy strategies that are inefficient but that reveal a high level of mathematical knowledge. This may be especially true for open-ended tasks with more complex responses. Although some bugs and misconceptions are frequent, students do not appear to apply them consistently. At the individual level, it will probably be most useful to address only the most common bugs and misconceptions, and only in situations where students show consistent patterns of responding. At the group or class level, it is sensible to address any bugs and misconceptions that occur frequently. Finally, research suggests that people can possess high levels of situation-specific competency; it may be useful to elicit evidence for this level, as this stage may be a stepping stone to more complete understanding.

## Prescriptions for the Design of Middle-School Mathematics Tasks

Now that mathematics competency in the middle grades has been characterized and the nature of evidence that is important to collect has been discussed, the features of tasks that will provide evidence for the identified competencies maybe considered. According to Mislevy, Steinberg, and Almond (2002), "A fundamental tenet of the evidenced-centered approach to assessment design (and of Messick's construct-centered approach as well) is that the characteristics of tasks are determined by the nature of the behaviors they must produce, to constitute evidence for the targeted aspects of proficiency" (p. 116). The content of the tasks developed should reflect the core content identified in the earlier section about characterizing mathematics competency in the middle grades. At the sixth grade, many of the

tasks should focus on number concepts and operations, and should be designed to provide a segue to the study of algebra. For seventh grade, there should be a heavy emphasis on proportional reasoning and algebraic concepts, and by eighth grade the tasks should require students to work with multiple expressions and graphs. In order to develop tasks that will elicit evidence for the competencies discussed in this report, it will be necessary to include a balance of tasks, some of which require complex responses and others that require only shorter responses to targeted prompts.

### *Complex Response Types*

Complex response types can take many forms. One type of complex response is an extended text response that presents an argument or justification. Complex responses are not necessarily lengthy, however, and they do not necessarily have to contain text. For example, proofs by dissection require a sequence of moves rather than explanatory text or expressions. As another example, a solution to an algebra problem might consist of a diagram followed by an ordered set of equations—and contain little if any text. In providing a complex response, students may invoke a variety of representations, depending on the affordances of the response type. In developing tasks that elicit complex responses, a wide variety of task types should be considered, and they may vary along any of the following dimensions: the nature of the prompt, the level of interaction, and the format of the responses.

Complex responses can provide valuable evidence about students' process competencies. For example, a complex response may consist of a mathematical argument that draws on tabular, graphic, and symbolic representations—the response may reflect strategic decisions as well as procedural fluency and background knowledge. As discussed earlier, however, often different strategies may be applied to the solution of a problem, and different strategies provide different evidence and require different amounts of time. From a fairness perspective, it is especially important to consider the impact of strategy selection if a nonjudicious strategy choice will compete with the amount of time a student has to complete the remaining tasks in an assessment. One possible solution is to provide students with some guidance regarding how to approach tasks that require complex responses. This is often done by breaking a task down into multiple parts, where responses to earlier parts constitute subgoals for the eventual solution, which is usually the goal of a later part. It is possible that even fairly subtle task modifications may help provide structure that could prevent students

from embarking on an overly complex solution procedure. Catrambone (1996) found that, when presenting students with worked examples, providing even noninformative labels for related sequences of steps helps them to establish subgoals during problem solving on subsequent transfer tasks.

Of course, imposing even subtle constraints may result in tasks that provide less evidence of strategic planning. In a formative assessment system, it is probably acceptable to impose few, if any, constraints because the results will never be used for accountability purposes. Teachers and students can use these tasks as an opportunity to compare and contrast different solution methods, including discussions of why they are equivalent and why some may be more efficient than others. In an accountability assessment system, however, strategy by task interactions raise concerns with respect to fairness in scoring. Particularly with tasks that require complex responses, it will be necessary to conduct both cognitive task analyses and pilot studies that focus on alternative strategies and how they differ, both in terms of the evidence they provide and the time they require.

Some complex response types provide an opportunity for students to formulate and develop mathematical arguments, an important process competency discussed earlier. Although there are many simple response types that assess mathematical reasoning skill, they do not provide evidence for how well a student can structure an argument or how clearly he or she is able to explain it. During the course of a complex response, a student may have the opportunity to communicate mathematical ideas with clarity, perhaps using a variety of representations. The extent to which this kind of thinking is in evidence should form the basis of the scoring rubric.

A prompt that requires a complex response must be very carefully constructed so that students do not misunderstand the intent. Marshall (1995b) described her experience as a member of the Mathematics Advisory Committee for the California Assessment Program (CAP). After the CAP field tested open-ended items with 12th-grade students, members of the advisory committee reviewed a sample of responses from among the large number collected—their findings were discussed in the document *A Question of Thinking*. Marshall discussed one of the items summarized in that report: about one-quarter of the students who answered this particular item did not respond in mathematical terms, focusing instead on situational factors. Her point was that many of the students may not have understood the

purpose of the task. This is not an uncommon experience in the development of open-ended problems, where there are no options to provide cues about what constitutes an acceptable response. Attention to the design of the prompt can help avoid this situation; breaking a task into parts can also help provide supporting structure that can clarify how a task should be interpreted.

There is a final reason why tasks that require extended responses may be helpful to include, in addition to the nature of evidence they provide. Such tasks may also support learning, because they require students to explain their reasoning. Students who provide higher quality self-explanations during the learning of worked examples tend to show higher levels of problem-solving performance (Chi, Bassok, Lewis, & Reimann, 1989; Pirolli & Recker, 1994). Also, prompting students to explain during the learning of worked examples improves subsequent performance (Chi, de Leeuw, Chiu, & LaVancher, 1994). However, eliciting self-explanations is not helpful under all circumstances or for all types of tasks. In situations where making self-explanations imposes additional cognitive load and competes with the resources needed for learning, they may not be of benefit (Nathan, Mertz, & Ryan, 1994). So while students often benefit from making explanations while studying worked examples, it is not clear that this will be of benefit when making explanations is part of the task. Further research should investigate the impact on learning of providing explanations during an assessment.

### *Basic Response Types*

Basic response types include multiple-choice and multiple selection–multiple choice, as well as some constructed response types, including: numeric entry, mathematical expressions and equations, some kinds of graphs, and short-answer text. Note that the term *basic* does not imply that these problems are easy to solve, only that an answer, if found, is straightforward to indicate, relative to a complex response type. These types also have an extremely important role to play in a middle-school mathematics assessment program designed to encourage learning. Basic responses may be useful for diagnosing misconceptions. As discussed earlier, since students are not necessarily consistent in how they endorse misconceptions, it is important not to overinterpret a single student response. It is certainly worthwhile, however, to develop prompts that target particular common misconceptions (such as the incorrect addition strategy). Once important misconceptions

have been identified, there is still the question of how to organize them in such a way that they may be meaningfully incorporated into tasks, and there are a number of approaches to this. Minstrell (2001) uses the *facet* as the unit of organization. Each content area is subdivided into several concept clusters, and each cluster describes a number of student facets, or ideas. In the design of a particular item, options (or possible responses) are developed in accordance with particular facets. The eventual response made by the student is identified with respect to the facet it represents.

Bart, Post, Behr, and Lesh (1994) developed the notion of a *semi-dense* item to guide the development of informative multiple-choice items. In order to qualify as semi-dense, an item has to meet the following requirements: each possible response must be interpretable in accordance with one and only one cognitive rule, and each relevant cognitive rule must be represented by a possible response. A cognitive rule could be a misconception, a procedural bug, or a correct idea. Although the semi-dense notion was developed to guide multiple-choice item construction, the idea may be easily extended to the development of constructed response types. Instead of constructing options that are in a one-to-one correspondence with cognitive rules, one considers the correspondences between *possible* responses and cognitive rules, and this can help guide the design of the prompt. For example, if more than one misconception could lead to the same generated response, one might reword the item (to use different numbers, perhaps) so that this scenario does not occur. In practice, it is challenging to develop items that meet the criteria for semi-density. Nevertheless, collectively these criteria provide a useful gold standard against which to compare any diagnostic tasks that are developed.

Cromley and Mislevy (2004) used a template-based approach to organize misconceptions for use in an assessment, incorporating the identification of misconceptions into an evidence-centered design approach.

Basic response types may also be useful in situations that assess recognition of mathematical structures. In some situations, it may be of greater interest to assess schema recognition than the details of solution. Experienced algebra problem solvers are able to quickly classify algebra word problems into schematic categories, and appear to use these categorizations during solution (Hinsley, Hayes, & Simon, 1977). Bennett, Sebrechts, and Rock (1995) piloted two categorization task types for the GRE program. In tasks of the first

type, examinees sorted algebra word problems by matching them to category exemplars. In tasks of the second type, examinees and experts rated the similarity of pairs of algebra word problems. For both types of tasks, examinee performance was positively correlated with admission test scores. On the similarity task, high-performing examinees made ratings that were more like experts'. Examinees preferred the sorting task to standard multiple-choice items, but preferred standard multiple-choice items to the similarity ratings task. Bennett et al. suggested that categorization tasks might also be used for diagnostic purposes, but that future research would need to determine whether student weaknesses may be specific to a particular aspect of problem solving (e.g., representation vs. solution). Categorization tasks might be candidates for inclusion in either a formative or accountability assessment.

Basic response types are also extremely useful when the purpose is to determine whether a student has knowledge of a specific fact or procedure. As discussed earlier, highly open-ended prompts that require complex responses may elicit the use of any of a number of strategies, and so the responses may provide different evidence, both with respect to background knowledge and procedural fluency. Complex response types are therefore not necessarily suited for assessing a student's knowledge of a particular fact or skill. It is important to make sure that such "knowledge and skill check" types are included; otherwise students may develop selective preferences for particular procedures or disregard specific facts.

Again, the importance of encouraging the development of procedural fluency should be emphasized. As described earlier, conceptual understanding and procedural fluency interact, and it would be a mistake to develop tasks that as a set only assess one or the other. It will be important to include tasks that encourage the development of procedural fluency as well as conceptual understanding. It has already been discussed how procedural tasks can reflect a lack of conceptual understanding. But developing students' skill with procedures may indirectly enhance conceptual understanding, as the following quote from Silver (1987) suggests:

> Students' problem-solving abilities might improve greatly if they could use working memory more efficiently, that is, if they learned to use automatic processing for the more routine elements of an activity, and thus made resources available for the controlled processing of the novel aspects of solving the assigned problems. (p. 40)

Another advantage of administering basic response types is that it is possible to automatically score them very accurately. Bennett, Morley, and Quardt (2000) discussed three constructed response types: mathematics expressions (ME), generating examples (GE), and graphical modeling (GM). The ME type requires a mathematics expression as a response, the GE type requires that the student provide an example that meets certain mathematical constraints, and the GM type requires the student to plot a function. These types place different cognitive demands on the student, but all of them have keys that can be expressed in mathematical terms and all can be automatically scored. At this stage, designing keys for automatic scoring of constructed response types can be challenging and requires specific technical knowledge, but several tools are under development at ETS that should make the key definition process both more accessible and more generalizable.

### *Prompt Complexity*

So far we have justified a need for both complex and basic response types. Complex response types that require an extended text response will usually take longer to complete than basic response types. But the expected length of the response is not the only feature that determines how a student will interact with a task element or how long it will take to complete. The task prompts may also vary in complexity. Some problems may require a good deal of reflection followed by a concise response. For example, in a data analysis and probability problem, it is important to assess whether a student can draw meaningful conclusions and notice trends from data presented in a table or graph. For this type of problem, the student may spend the majority of his or her time studying the data that is displayed and drawing inferences from it.

When administering exams to her graduate statistics classes, Marshall (1995b) provides students with statistical output and asks questions about the outputs. She does this so that students do not have to focus on statistical computations, at least for some of the tasks. Statistical output is an example of a highly complex prompt—more complex than would be encountered in a middle-school mathematics assessment—but the example will serve to illustrate the point. Marshall recognized that the outputs take time to absorb, and she didn't want students to spend exam time figuring out the outputs; nor did she want students who read more slowly to be pressed for time. She solved the problem by distributing the outputs to students a week before the test, so that they would have time to reflect on the

information contained in them. The students spend their exam time answering questions that refer to the outputs. This kind of solution would work very well in the context of a formative assessment, as would the approach of having students iteratively revise data models in a group setting (as in the work of Lehrer & Schauble, 2000, described earlier).

While allowing students to reflect on materials in advance might work very well in a formative setting, for security reasons it would not be advisable in an accountability setting. For this reason, it is likely that tasks developed for accountability assessments should not require students to read a lot of background text or to sift through material—if a problem is based on a large amount of data or information, it should be presented in consolidated or summarized form.

### *Cognitive Load and Task Design*

Complex tasks present a design challenge because they are difficult to develop without introducing construct-irrelevant features. Scenarios that are so complicated that they distract from the main goal should be avoided. Also, it may be that while the goal of some tasks is to find out what students know, the primary goal of other tasks may be to support learning. And it may be that the same tasks cannot serve both goals in all instances. Sweller (1992) provided a number of suggestions for reducing cognitive load in mathematics tasks designed to facilitate learning. One of his suggestions was to include goal-free tasks. His argument was that in solving goal-specific tasks, students often work backwards from the solution, using means-ends analysis. This requires cognitive resources that may compete with learning. The particular example he used was a geometry problem in which the student must find the value of a particular angle. This is an effortful process, requiring the student to work backwards from the unknown until he or she finds an angle to relate to the goal angle. The student must then retrace his or her steps to compute the value. In a goal-free problem, the student would be asked to find the values of all the angles that appear in the diagram. Sweller and colleagues have corroborated the finding that working with goal-free tasks improves learning in geometry, trigonometry, and kinematics (as cited in Sweller, 1992).

The potential benefit of learning from worked examples has also been documented (Sweller, 1992; Zhu & Simon, 1987). Most of the positive findings from the self-explanations literature discussed earlier involve explanations of worked examples. Sweller's account of why learning from worked examples is effective is again that studying examples

imposes less cognitive load than solving problems, giving students the opportunity to attend to features that support subsequent problem solving. To ensure that students have incentive to learn as much as possible from the worked examples, Sweller recommends alternating worked examples with tasks to be solved.

Finally, Sweller (1992) noted that students do not always learn successfully from worked examples. Although studying worked algebra problems was often helpful, students did not derive the same benefit from studying worked geometry examples, at least initially. The geometry examples initially consisted of a diagram followed by lines of text that described the steps in solving the problem. The text typically referred to angles in the diagram, forcing the student to shift back and forth between the diagram and the text in order to consolidate all the information in the problem. Sweller and colleagues found that by integrating the text into the diagram so that the problem-solving steps were near the angles they referred to made the geometry examples function more like the algebra examples. In other words, students derived a learning benefit from studying them. Again, Sweller explained this result in terms of reduced cognitive load.

What do Sweller's recommendations imply for the design of a middle-school mathematics assessment? There is no reason to include tasks that introduce cognitive load unnecessarily. Goal-free tasks may be useful to include in either formative or accountability assessments. One possibility for including goal-free tasks might be in a multipart task. An early part might ask the student to find values for many unknowns, and a later part might ask the student for a particular part (the goal). In a formative assessment, it may be very helpful to intersperse worked examples with problem-solving tasks. Providing worked examples might also be an opportunity to present examples in accordance with Michener's classification scheme, described earlier. Time constraints will probably not allow the use of worked examples in an accountability assessment.

### *Interactive Task Components*

Manipulatives have been used as part of mathematics instruction for years. More recently, software such as Geometer's Sketchpad is becoming widely used. Such tools provide students with a way to reason concretely about more abstract concepts and allow for experimentation. We should consider incorporating virtual manipulatives and simulations in computer-based assessments that are developed. In the development of formative

assessments, there is the potential for flexibility in how interactive task components are incorporated and used by students. The interaction in an accountability system will need to be limited and simple. Students should not be spending any time struggling with the interface or stuck at an impasse from which there is no easy return.

ETS has developed a graphical interface in which students can plot functions by clicking on a sequence of points. Since these items may be automatically scored, we may make heavy use of this capability for integrating interactive components. Spreadsheets may be incorporated into tasks that involve transformations of tabular data.

### *Using Item Modeling and Automatic Item Generation to Support Large-Scale Task Development and Formative Assessment*

An assessment system that consists of both an accountability component and a formative component will require classes of items, where each item in a class addresses a subset of a given constellation of competencies. Note that the term *item* here refers to any task element to which a student would respond. On the accountability side, if some skills will be reassessed across periodic administrations, multiple items from each class are needed in order to measure a student's standing with regard to identical and/or related skills at different time points. Since the formative materials are intended to support learning of the skills assessed by the accountability component, multiple items from each class are needed for both components. Item modeling is an approach that can support systematic development of classes of related items. Although it is unlikely that item modeling and automatic item generation can significantly support the development of complex, extended, and highly situated tasks that are included in the CBAL assessment system, it can support the development of shorter diagnostic tasks and may be able to support the development of parts of the more extended tasks.

LaDuca, Staples, Templeton, and Holzman (1986) used the term *item model* to refer to classes of items that assess the same content. Hively, Patterson, and Page (1968) referred to related mathematics items as *item forms*. As in Bejar (2002), we use the term *item model* to refer to a class of items that share a common set of characteristics. In particular, we focus on quantitative item models, which share specifications expressed in mathematical terms.

Assessment items may be organized into classes along a number of different dimensions. Most often, mathematics items are classified based on their underlying

quantitative structure—Mayer's (1981) taxonomy of algebra story problems is an example. Items may also be classified in accordance with other characteristics, such as numerical complexity, surface features, or item format.

Components of an item model that may change are represented as *variables*; *constraints* specify how variables are related. Item models may be represented by item shells that consist of static text interspersed with slots for the variables. This sort of representation is sufficient in many situations; however, it does have limitations. For further discussion on this point, see Higgins, Futagi, and Deane (2005) and Deane, Graf, Higgins, Futagi, and Lawless (2006).

An item model can be used as a guide to generate items by hand, but it may also be programmed with a computer, and items may be automatically generated from the program (e.g., Meisner, Luecht, & Reckase, 1993; Singley & Bennett, 2002). When an item is generated, random values that satisfy the constraints are assigned to each variable in the model. Each item generated from an item model is referred to as an *instance* (Bejar, 2002). Bejar described how item models may be designed to generate one of two types of instances: *isomorphs* or *variants*. Isomorphs share a common deep structure and have similar psychometric parameters; variants differ systematically in this regard. For example, an item model may be designed to generate some items that are easy and others that are difficult. Irvine (2002) referred to a variable that influences psychometric parameters as a *radical* and a variable that does not as an *incidental*.

Item models may be used to generate instances in different standard formats, including multiple-choice and constructed response. Variables may be incorporated into any part of an item model, including the stem, the key, and if applicable, the distractors. For example, a constructed response item model includes a *stem model* and a *key model*, and a multiple-choice item model includes a *stem model*, as well as *option models*. Item models may also be used to describe instances in more novel formats, including multipart questions that may be scored using a partial credit rubric. Finally, modeling may be extended to generate complex scoring keys, answer choice rationales, and response-specific feedback. For example, Morley, Lawless, and Bridgeman (2005) modeled answer choice rationales. Modeling response feedback so that it is customized to each instance can be accomplished by

reusing variables from the stem model or any of the option models in a *feedback model*. An example of this is shown in Graf, Steffen, Peterson, Saldivia, and Wang (2004).

Automatic item generation may be an economical approach to test development in large-scale assessment programs (Bejar et al., 2002). First, items are generated from an algorithm, rather than individually through a manual process. Second, the approach may save on pretesting costs. If the psychometric parameters of model-generated instances may be predicted successfully in advance, it may not be necessary to calibrate each instance individually (e.g., Bejar, 1993, 1996; Bejar et al., 2002; Bejar & Yocom, 1991; Embretson, 1999). Several researchers have explored the extent to which item models generate isomorphs, or instances with highly similar psychometric parameters (Meisner, Luecht, & Reckase, 1993; Sinharay & Johnson, 2005; Steffen, Graf, Levin, Robin, & Lu, 2006). The general result from this work is that some models generate instances with very similar parameters while others generate instances with highly variable parameters, and how a model will behave is not always clear at the outset. Because of this, an iterative approach to item model development, where empirical evaluation is followed by subsequent revision, is recommended (Bejar, 1993; Bejar & Yocom, 1991; Embretson & Gorin, 2001; Graf, Peterson, Steffen, & Lawless, 2005).

A model-based approach may also enhance construct validity, because it requires that the relationships between generative principles and psychometric properties be made explicit (Bejar, 1993; Bejar & Yocom, 1991). Item modeling lends itself to being used in conjunction with experimental designs that systematically explore features that influence item difficulty and discrimination (Bejar, 1993; Bejar & Yocom, 1991; Embretson, 1999; Enright, Morley, & Sheehan, 2002; Enright & Sheehan, 2002; Graf et al., 2005; Newstead, Bradon, Handley, Evans, & Dennis, 2002).

Although the most obvious application of item modeling may be to support development in large-scale testing programs, we should consider how to use a model-based approach to support the development of diagnostic, classroom-based assessments. These kinds of assessments could be used to guide instruction, as part of a formative approach. A formative approach to assessment has been shown to have a positive impact on student learning (Wiliam, Lee, Harrison, & Black, 2004).

The goals for developing diagnostic item models for formative assessment are quite different from the goals of the research described earlier. First, we are less concerned with generating instances with psychometric parameters that can be predicted very accurately, and more concerned with generating instances that consistently measure patterns of understanding with accuracy sufficient to focus instruction. We should also be exploring opportunities to automatically generate targeted response feedback and partial-credit scoring rubrics.

## Concluding Summary

The purpose of this report is to provide a set of recommendations to guide the design of formative and accountability assessments for middle-school mathematics. The document is divided into four main sections. The Mathematical Competency section discusses important aspects of mathematical competency; mathematical competency is characterized with respect to both core *content* and key *processes*. It concludes that in the middle grades, it is most important to focus on algebra and the connections between algebra and numbers and operations. This is not to suggest, however, that this focus will come at the expense of other core content that is emphasized in the middle grades. The competency models presented in the second section propose developmental trajectories for content (as inferred from the NCTM curriculum focal points) and process. On the process side, students should develop increasingly abstract representations to support increasingly sophisticated mathematical arguments.

The next section, Describing and Quantifying Evidence of Mathematical Proficiency, describes the features of responses that provide evidence with respect to the specified competencies. This section focuses on developmental progressions, strategies, bugs and misconceptions, and the role of the situative perspective. One of the concerns expressed in this section is how to contend with the challenge of strategy by task interactions. Open-ended response types in particular lend themselves to the use of alternate strategies. Different strategies may reveal very different evidence and take different amounts of time to complete. This is especially a concern in the development of accountability assessments. One possible solution is to provide just enough guidance in the structure of the prompts to ensure that students do not pursue unwieldy strategies or find themselves stuck at an unrecoverable impasse. The review of the literature on bugs and misconceptions led to the conclusion that it

is worthwhile to diagnose very common misconceptions, especially in situations where students show consistent patterns of responding. More often than not, however, students are not consistent in the bugs or misconceptions they endorse, so it is not cost-effective to try to identify and diagnose each possible bug or misconception.

The final section, Prescriptions for the Design of Middle-School Mathematics Tasks, makes recommendations for features of tasks that may also be considered learning events. Tasks that require extended responses and tasks that require short answers will both be needed in order to provide evidence of the specified competencies. Complex response types are ideal for eliciting evidence of mathematical processes, including how students use alternate representations and develop arguments. Since complex response types lend themselves to alternate solution strategies, students are not always required to draw on specific knowledge or procedures. The basic response types are better suited for identifying common misconceptions and for determining whether a student has knowledge of a particular skill or procedure. In an effort to minimize the influence of construct-irrelevant variables, tasks should not impose any unnecessary cognitive load. Finally, in order to support the development of assessments on a large scale, a model-based approach to task development is recommended, as is the continued enhancement of automatic scoring capabilities.

## References


American Association for the Advancement of Science. (2001). *Atlas of science literacy.* Washington, DC: Author.

Anderson, J. R., Greeno, J. G., Reder, L. M., & Simon, H. A. (2000). Perspectives on learning, thinking, and activity. *Educational Researcher, 29*(4), 11-13.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher, 25*(4), 5-11.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1997). Situative versus cognitive perspectives: Form versus substance. *Educational Researcher, 26*(1), 18-21.

Angeles, M., Sampang, A., & Moseros, J. (2006, May). *Redesigning the CEM mathematics diagnostic tests as developmental assessment instruments*. Paper presented at the IAEA 2006 annual conference, Singapore.

Bart, W. M., Post, T., Behr, M. J., & Lesh, R. (1994). A diagnostic analysis of a proportional reasoning test item: An introduction to the properties of a semi-dense item. *Focus on learning problems in mathematics,* 16(3), 1-11.

Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement, 21*(2), 175-189.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen (Ed.), *Test theory for a new generation of tests* (pp. 323-357). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (ETS Research Rep. No. RR-96-13). Princeton, NJ: ETS.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-218). Mahwah, NJ: Lawrence Erlbaum Associates.

Bejar, I. I., Graf, E. A., & Oranje, A. (2008). *Form models as an approach to achieving score comparability.* Manuscript submitted for publication.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). *A feasibility study of on-the-fly item generation in adaptive testing* (ETS Research Rep. No. RR-02-23). Princeton, NJ: ETS.

Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement, 15*(2), 129-137.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43-61). New York: Springer.

Bennett, R. E., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement, 24*(4), 294-309.

Bennett, R. E., Sebrechts, M. M., & Rock, D. A. (1995). *A task type for measuring the representational component of quantitative proficiency* (ETS Research Rep. No. RR-95-19). Princeton, NJ: ETS.

Bernardo, A. B. I., & Okagaki, L. (1994). Roles of symbolic knowledge and problem-information context in solving word problems. *Journal of Educational Psychology, 86*(2), 212-220.

Birenbaum, M., Tatsuoka, C., & Yamada, T. (2004). Diagnostic assessment in TIMSS-R: Between-countries and within-country comparisons of eighth graders' mathematics performance. *Studies in Educational Evaluation, 30*, 151-173.

Borchert, K. (2000). *Connecting operational and structural knowledge in algebra: The impact of word problem solving on equation construction*. Unpublished master's thesis, University of Washington, Seattle, WA.

Borchert, K. (2003). *Dissociation between arithmetic and algebraic knowledge in mathematical modeling*. Unpublished doctoral dissertation, University of Washington, Seattle, WA.

Briars, D. J., & Larkin, J. H. (1984). An integrated model of skill in solving elementary word problems. *Cognition & Instruction, 1*(3), 245-296.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*(2), 155-192.

Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education, 15*(3), 179-202.

Carraher, T. N., Carraher, D. W., & Schliemann, A. D. (1985). Mathematics in the streets and in schools. *British Journal of Developmental Psychology, 3*, 21-29.

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(4), 1020-1031.

Chi, M. T., Bassok, M., Lewis, M. W., & Reimann, P. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145-182.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Clement, J., Lochhead, J., & Monk, G. S. (1981). Translation difficulties in learning mathematics. *American Mathematical Monthly, 88*(4), 286-290.

Clement, J., Lochhead, J., & Soloway, E. (1979). *Translating between symbol systems: Isolating a common difficulty in solving algebra word problems.* Amherst: University of Massachusetts.

Cromley, J. G., & Mislevy, R. J. (2004). *Task templates based on misconception research* (No. CSE Report 646). College Park, MD: University of Maryland.

Deane, P., Graf, E. A., Higgins, D., Futagi, Y., & Lawless, R. (2006). *Model analysis and model creation: Capturing the task-model structure of quantitative item domains* (ETS Research Rep. No. RR-06-11). Princeton, NJ: ETS.

Donovan, M. S., & Bransford, J. D. (Eds.). (2005). *How students learn: Mathematics in the classroom. Committee on how people learn, a targeted report for teachers*. Washington, D.C: National Academies Press.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*(4), 407-433.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343-368.

English, L. D., & Halford, G. S. (1995). *Mathematics education: Models and processes*. Hillsdale, NJ: Lawrence Erlbaum.

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15*(1), 49-74.

Enright, M. K., & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: Implications for item generation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum Associates.

Epp, S. (2003). The role of logic in teaching proof. *American Mathematical Monthly, 110*(10), 886–899.

Falmagne, J.-C., Cosyn, E., Doignon, J.-P., & Thiéry, N. (2006). The assessment of knowledge, in theory and in practice. In R. Missaoui & J. Schmid (Eds.), *Lecture notes in artificial intelligence, Vol. 3874. ICFCA 2006* (pp. 61-79). Berlin: Springer-Verlag.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1993). *The effects of statistical training on thinking about everyday problems.* In R. E. Nisbett (Ed.), *Rules for reasoning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gitomer, D. H., & Steinberg, L. (1999). Representational issues in assessment design. In I. C. Sigel (Ed.), *Development of mental representation: Theories and application*. Princeton: ETS.

Goldin, G. A. (1998). Representational systems, learning, and problem solving in mathematics. *Journal of Mathematical Behavior, 17*(2), 137–165.

Graf, E. A., Bassok, M., Hunt, E., & Minstrell, J. (2004). A computer-based tutorial for algebraic representation: The effects of scaffolding on performance during the tutorial and on a transfer task. *Technology, Instruction, Cognition, and Learning, 2*(1-2), 135-170.

Graf, E. A., Harris, K., Marquez, E., Fife, J., & Redman, M. (2009). *Cognitively based assessment of, for, and as Learning (CBAL) in mathematics: A design and first steps toward* implementation (Research Memorandum No. RM-09-07). Princeton, NJ: ETS.

Graf, E. A., Peterson, S., Steffen, M., & Lawless, R. (2005). *Psychometric and cognitive analysis as a basis for the design and revision of quantitative item mode* (ETS Research Rep. No. RR-05-25). Princeton, NJ: ETS.

Graf, E. A., Steffen, M., Peterson, S., Saldivia, L., & Wang, S. (2004, October). *Designing and revising quantitative item models.* Presentation at the 4th Spearman conference, Princeton, NJ.

Greeno, J. G. (1997). On claims that answer the wrong questions. *Educational Researcher, 26*(1), 5-17.

Greeno, J. G., Brown, J. S., Foss, C., Shalin, V., Bee, N. V., Lewis, M. W., et al. (1986). *Cognitive principles of problem solving and instruction. Final report*. Pittsburgh, PA: Learning Research and Development Center.

Hale, M. (2002). *The tree of mathematics*. Retrieved December 23, 2008, from http://www.stetson.edu/~mhale/logic/tree.htm

Hall, R., Kibler, D., Wenger, E., & Truxaw, C. (1989). Exploring the episodic structure of algebra story problem solving. *Cognition and Instruction, 6*, 223-283.

Hart, K. (1984). *Ratio: Children's strategies and errors.* Windsor, England: NFER-Nelson.

Higgins, D., Futagi, Y., & Deane, P. (2005). *Multilingual generalization of the ModelCreator software for math item generation.* (ETS Research Rep. No. RR-05-02). Princeton, NJ: ETS.

Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations meaning and representation in algebra word problems. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 89-106). Oxford, England: Lawrence Erlbaum Associates.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5*(4), 275-290.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: The MIT Press.

Hoyles, C., Noss, R., & Pozzi, S. (2001). Proportional reasoning in nursing practice. *JRME Online, 32*(1), 1-38.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 51-74). Cambridge, MA: The MIT Press.

Irvine, S. H. (2002). The foundations of item generation for mass testing. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 3-34). Mahwah, NJ: Lawrence Erlbaum Associates.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237-251.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11*(2), 143-157.

Kaput, J. J. (1999). Teaching and learning a new algebra with understanding. In E. Fennema & T. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 133-155). Mahwah, NJ: Lawrence Erlbaum Associates.

Kaput, J. J., & West, M. (1994). Missing-value proportional reasoning problems: Factors affecting informal reasoning patterns. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 235-287). New York: State University of New York Press.

Karplus, R., Pulos, S., & Stage, E. K. (1983). Proportional reasoning of early adolescents. In R. Lash & M. Landau (Eds.), *Acquisition of mathematics concepts and processing* (pp. 45-86). New York: Academic Press.

Kieran, C. (1992). The learning and teaching of school algebra. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 390-419). New York : Macmillan Publishing Co Inc.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Koedinger, K. R., & Anderson, J. R. (1998). Illustrating principled design: The early evolution of a cognitive tutor for algebra symbolization. *Interactive Learning Environments, 5*, 161-179.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity? *Intelligence, 14*(4), 389-433.

LaDuca, A., Staples, W. I., Templeton, B., & Holzman, G. B. (1986). Item modelling procedure for constructing content-equivalent multiple choice questions. *Medical Education, 20*, 53-56.

Lave, J., Murtaugh, M., & de la Rocha, O. (1984). The dialectic of arithmetic in grocery shopping. In J. Lave & B. Rogoff (Eds.), *Everyday cognition: Its development in social context* (pp. 67-94). Cambridge, MA: Harvard University Press.

Lee, L., & Wheeler, D. (1989). The arithmetic connection. *Educational Studies in Mathematics, 20*, 41-54.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1993). The effects of graduate training on reasoning : Formal discipline and thinking about everyday-life events. In R. E. Nisbett (Ed.), *Rules for reasoning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (pp. 101-159). Mahwah, NJ: Lawrence Erlbaum Associates.

Lesh, R., & Lamon, S. J. (1992). Assessing authentic mathematical performance. In *Assessment of authentic performance in school mathematics* (pp. 17-57). Washington, DC: American Association for the Advancement of Science.

Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology, 79*(4), 363-371.

Madhyastha, T., Hunt, E., Kraus, P., & Minstrell, J. (2006). *The relationship of coherence of thought and conceptual change to ability.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Marshall, S. P. (1995a). *Schemas in problem solving*. New York: Cambridge University Press.

Marshall, S. P. (1995b). Some suggestions for alternative assessments. In S. F. Chipman, P. D. Nichols, et al. (Eds.), *Cognitively diagnostic assessment* (pp. 431-453). Hillsdale, NJ: Lawrence Erlbaum Associates.

Martin, S. A., & Bassok, M. (2005). Effects of semantic cues on mathematical modeling: Evidence from word-problem solving and equation construction tasks. *Memory and Cognition, 33*(3), 471-478.

Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science, 10*, 135-175.

Mayer, R. E. (1983). *Thinking, problem-solving, cognition*. New York: W. H. Freeman & Company.

Mayer, R. E., Larkin, J. H., & Kadane, J. B. (1984). A cognitive analysis of mathematical problem-solving ability. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 231-273). Hillsdale, NJ: Lawrence Erlbaum Associates.

Meisner, R., Luecht, R., & Reckase, M. D. (1993). *The comparability of the statistical characteristics of test items generated by computer algorithms* (ACT Research Rep. No. 93-9). Iowa City, IA: American College Testing.

Michener, E. R. (1978). Understanding understanding mathematics. *Cognitive Science, 2*, 361-383.

Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In K. Crowley (Ed.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 415-443). Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 97-128). Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives, 1*, 3-67.

Morales, R. V., Shute, V. J., & Pellegrino, J. W. (1985). Developmental differences in understanding and solving simple mathematics word problems. *Cognition and Instruction, 2*(1), 41-57.

Morley, M. E., Lawless, R. R., & Bridgeman, B. (2005). Transfer between variants of mathematics test questions. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 313-336). Greenwich, CT: Information Age Publishing.

Nathan, M. J., Mertz, K., & Ryan, R. (1994, April). *Learning through self-explanation of mathematics examples: Effects of cognitive load.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans. [

National Assessment Governing Board. (Ed.). (2005). *Mathematics framework for the 2005 national assessment of educational progress*. Washington, DC: Author.

National Assessment Governing Board. (2007). *Mathematics framework for the 2007 national assessment of educational progress*. Washington, DC: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics*. Retrieved September 12, 2006, from http://www.nctmmedia.org/cfp/full_document.

National Library of Virtual Manipulatives. (1999). Available from the Utah State University Web site: http://nlvm.usu.edu/en/nav/vLibrary.html.

National Research Council. (2000). *Mathematics education in the middle grades: Teaching to meet the needs of middle grades learners and to maintain high expectations: Proceedings of a national convocation and action conference*. Washington, DC: National Academy Press.

Newstead, S., Bradon, P., Handley, S., Evans, J., & Dennis, I. (2002). Using the psychology of reasoning to predict the difficulty of analytical reasoning problems. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 35-51). Mahwah, NJ: Lawrence Erlbaum Associates.

Nickerson, R. S., Perkins, D. N., & Smith, E. E. (1985). *The teaching of thinking*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1993). Teaching reasoning. In R. E. Nisbett (Ed.), *Rules for reasoning* (pp. 297-314). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1993). The use of statistical heuristics in everyday inductive reasoning. In R. E. Nisbett (Ed.), *Rules for reasoning* (pp. 15-54). Hillsdale, NJ: Lawrence Erlbaum Associates.

Noelting, G. (1980). The development of proportional reasoning and the ratio concept: Part I – differentiation of stages. *Educational Studies in Mathematics, 11*(2), 217-253.

Payne, S. J., & Squibb, H. R. (1990). Algebra mal-rules and cognitive accounts of error. *Cognitive Science, 14*(3), 641-642.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Pirolli, P., & Recker, M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction, 12*(3), 235-275.

Polya, G. (1957). *How to solve it* (2nd ed.). Princeton, NJ: Princeton University Press.

RAND Mathematics Study Panel, & Ball, D. L. (2003). *Mathematical proficiency for all students: Toward a strategic research and development program in mathematics education* (No. MR-1643.0-OERI). Santa Monica, CA: The Office of Education Research and Improvement.

Resnick, L. B., Cauzinille-Marmeche, E., & Mathieu, J. (1987). Understanding algebra. In J. A. Sloboda & D. Rogers (Eds.), *Cognitive processes in mathematics* (pp. 169-203). New York: Clarendon Press.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153-196). Rochester, NY: Academic Press.

Rosnick, P., & Clement, J. (1980). Learning without understanding: The effect of tutoring strategies on algebra misconceptions. *Journal of Mathematical Behavior, 3*(1), 3-27.

Schliemann, A. D., & Nunes, T. (1990). A situated schema of proportionality. *British Journal of Developmental Psychology, 8*, 259-268.

Schoenfeld, A. H. (1987). What's all the fuss about metacognition? In A. H. Schoenfeld (Ed.), *Cognitive science and mathematics education* (pp. 189-215). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schoenfeld, A. H. (1994). Reflections on doing and teaching mathematics. In A. H. Schoenfeld (Ed.), *Mathematical thinking and problem solving* (pp. 53-70). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schoenfeld, A. H. (2006). What doesn't work: The challenge and failure of the What Works Clearing House to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher, 35*(2), 13-21.

Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics, 22*(1), 1-36.

Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, present, and future. In D. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 570-600). New York: Macmillan.

Siegler, R. S. (2003). Implications of cognitive science research for mathematics education. In J. Kilpatrick, W. B. Martin & D. E. Schifter (Eds.), *A research companion to*

*principles and standards for school mathematics* (pp. 219-233). Reston, VA: National Council of Teachers of Mathematics.

Sigel, I. E. (1999). Approaches to representation as a psychological construct: A treatise in diversity. In I. E. Sigel (Ed.), *Development of mental representation: Theories and applications* (pp. 3-12). Mahwah, NJ: Lawrence Erlbaum Associates.

Silver, E. A. (1987). Foundations of cognitive theory and research for mathematics problem-solving instruction. In A. H. Schoenfeld (Ed.), *Cognitive science and mathematics education: An overview* (pp. 33-60). Hillsdale, NJ: Lawrence Erlbaum Associates.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine (Ed.), *Item generation for test development* (pp. 361-384). Mahwah, NJ: Lawrence Erlbaum Associates.

Sinharay, S., & Johnson, M. (2005). *Analysis of data from an admissions test with item models* (ETS Research Rep. No. RR-05-06). Princeton, NJ: ETS.

Sleeman, D. (1984). An attempt to understand students' understanding of basic algebra. *Cognitive Science, 8*, 387-412.

Steffen, M., Graf, E. A., Levin, J., Robin, F., & Lu, T. (2006). *An investigation of the psychometric equivalence of quantitative isomorphs: Phase 1.*Unpublished manuscript.

Stevens, R., & Thadani, V. (2006, October). *A value-based approach for quantifying scientific problem solving effectiveness within and across educational systems*. Paper presented at the Maryland Assessment Research Center for Education Success [MARCES] conference, Assessing and Modeling Cognitive Development in School: Intellectual Growth and Standard Setting, University of Maryland, College Park, MD.

Sweller, J. (1992). Cognitive theories and their implications for mathematics instruction. In G. C. Leder (Ed.), *Assessment and learning of mathematics* (pp. 46-61). Hawthorn, Australia: Australian Council for Educational Research Ltd.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*(4), 901-926.

Tirre, W. C., & Pena, C. M. (1993). Components of quantitative reasoning: General and group ability factors. *Intelligence, 17*, 501-521.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105-110.

Van Lehn, K. (1983). On the representation of procedures in repair theory. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 197-252). Rochester, NY: Academic Press.

Vennebush, P., Marquez, E., & Larsen, J. (2005). Embedding algebraic thinking throughout the mathematics curriculum. *Mathematics Teaching in the Middle School*, *11*(2), 86-93.

Vergnaud, G. (1983). Multiplicative structures. In R. Lesh & M. Landau (Eds.), *Acquisition of mathematics concepts and processes* (pp. 127-174). New York: Academic Press.

Weber, K. (2003, June). Students' difficulty with proof. In A. Selden & J. Selden (Eds.), *Research Sampler, 8.* Retrieved January 20, 2009 from the Mathematical Association of America Web site: http://www.maa.org/t_and_l/sampler/rs_8.html

Weisstein, E. W. (2006). *Pythagorean theorem from MathWorld—A Wolfram Web resource*. Retrieved July 4, 2006, from http://mathworld.wolfram.com/PythagoreanTheorem.html

Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F. K. Lester, Jr. (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053–1098). Greenwich, CT: Information Age Publishing.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11*(1), 49-65.

Wollman, W. (1983). Determining the sources of error in a translation from sentence to equation. *Journal for Research in Mathematics Education, 14*(3), 169-181.

Zhu, X., & Simon, H. A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4*(3), 137-166.

## Notes

[1]It is an open question how an accurate picture of student performance for the year will be established, since student competency is likely to change over the course of the year, particularly if the formative component is effective. Bejar, Graf, and Oranje (2009) discussed three possibilities. One possibility is to provide students with alternate forms that reassess competencies assessed earlier in the year. A second possibility is to design tasks that are cumulative with respect to the knowledge and skills they require—tasks administered early in the year would assess more basic competencies, while tasks administered later in the year would assess more advanced competencies as well as their basic building blocks. A third possibility is to assign greater weight to PAAs administered later in the year.

[2] Only the headings of the first three objectives are quoted; for complete text descriptions of all the cognitive objectives, the reader should refer to Lesh and Lamon (1992).