An Analysis of Higher-Order Thinking on Algebra I End-of Course Tests

Tony Thompson
Department of Mathematics, Science, and Instructional Technology Education
College of Education, East Carolina University, Greenville, NC 27858
(252) 328-9358; thompsonan@ecu.edu

ABSTRACT. This research provides insight into one US state’s effort to incorporate higher-order thinking on its Algebra I End-of-Course tests. To facilitate the inclusion of higher-order thinking, the state used Dimensions of Thinking (Marzano et al., 1988) and Bloom’s Taxonomy (Bloom et al., 1956). An analysis of Algebra I test items found that the state’s initial interpretation and application of Dimensions of Thinking and Bloom’s Taxonomy was faulty and inconsistent; as a result, few Algebra I test items from 1998 and 2001 were found to assess higher-order thinking. Algebra I test items written in 2007 were found to be more cognitively complex. This case study provides several findings that other test item writers (whether for the classroom or large-scale assessments) can apply to their own situation regarding assessing for higher-order thinking. These includes the importance of: (a) considering the “familiarity” a student has with a test item in deciding whether the item is lower- or higher-order thinking; (b) using a mathematics specific assessment framework with a limited number of categories; and (c) not confusing test items placed in a “real-world” context with higher-order thinking.

Keywords: higher-order thinking, large-scale assessment, algebra, Bloom’s Taxonomy, Dimensions of Thinking, assessment frameworks
An analysis of higher-order thinking on Algebra I End-of Course tests

The ability to think at higher levels is considered a major instructional goal of education and a driving force behind efforts to reform mathematics education over the past twenty years (Costa, 2001; De Bono, 1994; Donovan & Bransford, 2005; Hiebert et al., 1997; Kulm, 1990; Lubezky, Dori, & Zoller, 2004; National Council of Teachers of Mathematics, 1989; 2000; National Research Council, 2001; Resnick, 1987; Stein, Grove, & Henningsen, 1996). However, teaching for higher-order thinking (HOT) is difficult, and some US educators are concerned that state testing makes teaching for HOT even more challenging (Barksdale-Ladd & Thomas, 2000; Jones et al., 1999; Jones, Jones, & Hargrove, 2003; Kohn, 2000; McNeil, 2000; Neill, 2003; Ravitch, 2010; Smith, 1991; Smith & Rottenberg, 1991). Their argument is that many US state exams primarily focus on lower-order thinking (LOT) (e.g., procedural skills; symbol manipulation) at the expense of HOT (e.g., problem solving; reasoning).

According to the National Center for Education Statistics (1996), teaching for HOT along with professional development in HOT were found to be two of the top five variables positively associated with improved student achievement. Students of teachers who teach for both LOT and HOT outperform students whose teachers only teach for LOT (Wenglinsky, 2002). However, national and international assessments in mathematics (e.g., Trends in Mathematics and Science Study [TIMSS], Programme for International Student Assessment [PISA], and National Assessment of Educational Progress [NAEP]) indicate that the US educational system is not preparing students to solve complex problems, or in general, to think at higher levels (Baldi, Jin, Green, Herget, & Xie, 2007; Kloosterman, 2010; Lee, Grigg, Dion, 2007; Mullis, Martin, & Foy, 2008). The poor performance of US students on national and international exams reflects research findings which indicate that most US teachers find it difficult to teach and assess for
HOT (Kulm, 1990; Senk, Beckman, & Thompson, 1997, Silver et al., 2009). However, the ability to engage in high-level reasoning in mathematics is necessary for the 21st century workplace including the development of future scientists, engineers, and mathematicians (Partnership for 21st Century Skills, 2008).

Many US states claim to have created standards that are challenging and meet the goal of teaching for HOT; however, corresponding state exams designed to measure these standards often assess mostly lower levels of thinking (Lane, 2004; Nichols & Sugrue, 1999; Resnick, Rothman, Slattery, & Vranek, 2003; Webb, 1999; 2002; 2007). As Resnick et al. (2003) noted regarding their analysis of several US state exams, “it is … easier to test the simplest knowledge and skills … leaving complex concepts, extending reasoning, and other higher-level cognitive demands underrepresented on the test” (p. 6).

US state exams help define what content is important and have been shown to play a role in influencing teachers’ classroom practices (Pedulla, Abrams, Madaus, Russell, Ramos & Miao, 2003). In many cases, teachers will often write their own tests to reflect the specific content and format of their state tests (Corbett & Wilson, 1991; Madaus, West, Harmon, Lomax, & Viator, 1992; McMillan, Myran, & Workman 1999; Smith, 1991). It is within this context that some educators propose that if state tests can successfully focus on HOT, then it will encourage and assist teachers to teach for HOT in their classrooms (Kulm, 1990; Yeh, 2001; Popham, 1987; Tankersley, 2007; Wiggins, 1993). From this point of view, the development of state examinations that assess higher-order thinking is key to facilitating the development of HOT by all students (Zoller & Tsaparlis, 1997). There is evidence to support this view; for those states that have exams that include problem solving and open-ended tasks (i.e., HOT), teachers report
using more tasks in the classroom similar to those on the test (Pedulla et al., 2003; Stecher, Barron, Chun, & Ross, 2000).

Since large-scale assessments (such as state exams) play an important role in what teachers teach and students learn, it is important to take a critical look at US state testing programs. (Clarke et al., 2003; Lane, 2004, Webb, 1999; 2002; 2007) In particular, it is important to ascertain whether or not the cognitive skills they claim to assess are actually included on the exams (McMillan, Myran, & Workman, 1999). Since the 1990s, one US state, North Carolina, has used the thinking skills from *Dimensions of Thinking* (Marzano et al., 1988) as a framework to develop items for its End-of-Course (EOC) tests (North Carolina Department of Public Instruction [NC DPI], 1999). The NC DPI (1996) chose to classify test items using *Dimensions of Thinking* because it allowed them to better assess:

… the mastery of higher level skills. The term “higher level skills” refers to the thinking and problem solving strategies that enable people to access, sort, and digest enormous amounts of information. It refers to the skills required to solve complex problems and to make informed choices and decisions. (p. 1)

More specifically, *Dimensions of Thinking* was used on the mathematics EOC tests to help:

… focus on problem solving and assesses a student’s ability to apply mathematical principles, solve problems, and explain mathematical processes. Problems are typically posed as real situations that students may have encountered, but are not always simple in nature (e.g., quadratic or exponential functions in science). (NC DPI, 1998, p. 4)

The NC DPI made two modifications to *Dimensions of Thinking*. First, *Dimensions of Thinking* has eight thinking skills: focusing, information gathering, remembering, organizing, analyzing, generating, integrating, and evaluating. The NC DPI collapsed the first three categories – focusing, information gathering, and remembering – to a category called knowledge. Second, the NC DPI added a thinking skill called applying not found in *Dimensions of Thinking*. 
According to NC DPI officials, *applying* is defined to be consistent with *application* in *Bloom’s Taxonomy* (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956).

Below is a summary of the modified version of *Dimensions of Thinking* thinking skills that the NC DPI (1999) used to classify items in its EOC tests:

- **Knowledge**: using focusing skills, information-gathering skills, and remembering skills.
- **Organizing**: arranging information so it can be used effectively; noting similarities and differences between or among entities; grouping and labeling entities on the basis of their attributes; sequencing entities according to a given criteria; changing the form but not the substance of information.
- **Applying**: demonstrating prior knowledge within a new situation. The task is to bring together the appropriate information, generalizations or principles that are required to solve a problem.
- **Analyzing**: clarifying existing information by examining parts and relationships; identifying attributes and components; identifying relationships and patterns; identifying errors.
- **Generating**: producing new information, meaning, or ideas, going beyond available information to identify what reasonably may be true; anticipating next events, or the outcome of a situation; explaining by adding details, examples, or other relevant information
- **Integrating**: combining information efficiently into a cohesive statement and changing existing knowledge structures to incorporate new information.
- **Evaluating**: assessing the reasonableness and quality of ideas; creating standards for making judgments; confirming the accuracy of claims.” (p. 2)

According to an NC DPI mathematics specialist, the thinking skills of *knowledge, organizing* and *applying* are considered LOT while *analyzing, generating, integrating,* and *evaluating* are considered HOT.

**Purpose of the study**

It is important to investigate the extent to which assessment items on US state exams are eliciting HOT (Lane, 2004; Messick, 1989; Tankersley, 2007). In their research on US state assessments, Webb (1999; 2002; 2007) and Resnick et al. (2003) found that some states do a better job than others in assessing HOT. However, their research reports on these state
assessments provided anonymity to the states and specific test items are not shared. This is due to their decision to analyze secured state exams that were in use at the time of their research. However, this prevented the researchers from sharing test items to use as examples in order to support their conclusions. Keeping states anonymous and not including test items in research studies makes it difficult for the education community and public-at-large to independently verify the conclusions or to gain deeper insight into testing for HOT.

The purpose of this research was to investigate how one US state, North Carolina, used Dimensions of Thinking and Bloom’s Taxonomy to classify items for its Algebra I EOC tests in 1998, 2001, and 2007. In the US, Algebra I is generally taken by students in the 8th or 9th grade. Approaching this research from an historical perspective allows the researcher to use released items from Algebra I EOC tests for analysis that can be shared with readers. This provides others the opportunity to verify or challenge the conclusions presented in this research. The specific research questions guiding this study were:

(1) How did the North Carolina Department of Public Instruction (NC DPI) use Dimensions of Thinking and Bloom’s Taxonomy to categorize thinking skills for Algebra I End-of-Course (EOC) test items in 1998, 2001 and 2007?

(2) Did the NC DPI’s use of Dimensions of Thinking and Bloom’s Taxonomy result in the creation of Algebra I EOC test items that were likely to assess higher-order thinking?

Defining higher-order thinking

Thinking skills resist precise forms of definition, but as Resnick (1987) and Marzano et al. (1988) noted, lower-order thinking (LOT) and higher-order thinking (HOT) can be recognized when each occurs. In the research literature, LOT is often characterized as the recall of information or the application of concepts or knowledge to familiar situations and contexts. Schmalz (1973) noted that LOT tasks requires a student “… to recall a fact, perform a simple operation, or solve a familiar type of problem; it does not require the student to work outside the
familiar” (p. 619). Senk et al. (1997) characterized LOT as solving tasks where the solution requires applying a well-known algorithm, often with no justification, explanation, or proof required, and where only a single correct answer is possible. In general, LOT is characterized as solving tasks while working in familiar situations and contexts; or, applying well-known algorithms familiar to the students.

In contrast, Resnick (1987) characterized HOT as requiring nuanced judgment and being non-algorithmic; i.e., where the “path of action … [is] not fully specified in advance” (p. 44). Similarly, Stein and Lane (1996) describe HOT as “the use of complex, non-algorithmic thinking to solve a task in which there is not a predictable, well-rehearsed approach or pathway explicitly suggested by the task, task instruction, or a worked out example” (p. 58). Senk et al. (1997) characterized HOT as solving tasks where no algorithm has been taught, where justification or explanation are required, and where more than one solution may be possible. The National Council of Teachers of Mathematics (NCTM, 1989) characterizes higher-order thinking as solving a non-routine problem. That is, a problem involving “a situation in which, for the individual or group concerned, one or more appropriate solutions have yet to be developed” (p. 10). In mathematics, what distinguishes a non-routine problem from a routine problem (or exercise) is whether the student already knows an algorithm that, when applied, will lead to a solution (Charles & Lester, 1982; Kantowski, 1977). In general, HOT involves solving tasks where an algorithm has not been taught or using known algorithms while working in unfamiliar contexts or situations.

Despite the challenge of defining HOT, this research required a definition of HOT in order to evaluate Algebra I EOC test items. The main differences between LOT and HOT as described in the literature is the familiarity a student has to a task and whether the student already knows an
algorithm or solution strategy that, when used properly, will lead to a correct solution. Therefore, for this research project, the definition of Stein and Lane (1996) was used since it is specific to mathematics and encapsulates most descriptions of HOT found in the literature: A test item was identified by the researcher as requiring higher-order thinking if—from the students’ perspective—it requires “the use of complex, non-algorithmic thinking to solve a task in which there is not a predictable, well-rehearsed approach or pathway explicitly suggested by the task, task instruction, or a worked out example” (p. 58). It is important to note that the definition of HOT used by this researcher is not the definition of HOT used by NC DPI in the creation of test items. According to NC DPI officials, HOT is characterized as the thinking skills of analyzing, integrating, generating, and evaluating as defined in Dimensions of Thinking.

**Methods**

Many of the test items used in this research along with the corresponding thinking skills each item was designed to measure are available to the public on the NC DPI website ([http://www.ncpublicschools.org/accountability/testing/eoc/](http://www.ncpublicschools.org/accountability/testing/eoc/)). Seventy test items based on the 1998 Standard Course of Study (for academic year 2000-2001) and 72 test items from the 2003 Standard Course of Study (for academic year 2006-2007) were available to be downloaded for analysis. There were no 1998 Algebra I EOC test items (based on the 1993 Standard Course of Study) available online; however, the researcher obtained two 1998 Algebra I End-of-Course tests (Forms S and T) from NC DPI. Each 1998 EOC test consisted of 81 items for a total of 304 Algebra I EOC test items.

For inter-rater reliability, a doctoral graduate assistant (who was a former secondary mathematics teacher) also classified test items. To assist in categorizing test items the following materials were used as references:
The modified *Dimensions of Thinking* framework materials published by NC DPI (1999)

*Dimensions of Thinking* (Marzano et al., 1988)

*Taxonomy of educational objectives: Part I, cognitive domain* (Bloom et al., 1956)

*Handbook on formative and summative evaluation of student learning* (Bloom, Hasting, & Madaus, 1971)

Bloom’s books were also used since the NC DPI framework included *applying* based on *application* in *Bloom’s Taxonomy*. In addition, textbooks used by Algebra I teachers in North Carolina were consulted to determine the types of tasks students were likely to have been taught during Algebra I.

In March 2010, all 304 Algebra I test items from 1998, 2001, and 2007 were categorized using the definitions of LOT and HOT (Stein & Lane, 1996). The inter-rater reliability for categorizing the test items was 89%. According to Kaid and Wadsworth (1989), researchers can generally be satisfied with inter-reliability greater than 85%. Items for which the raters disagreed were resolved in NC DPI’s favor. In May 2010, all 304 Algebra I EOC test items from 1998, 2001, and 2007 were organized by thinking skill as classified by NC DPI. Similarities and differences within the classifications were identified to ascertain how the NC DPI applied *Dimensions of Thinking* and *Bloom’s Taxonomy* for Algebra I EOC test items. NC DPI’s categorization of thinking skills was analyzed and summarized using descriptive statistics.

It is important to note that 162 of the test items came from the two 1998 Algebra I EOC tests. The test items from 2001 and 2007 were sample test items and not in the form of an EOC test. However, test items on a released 2009 Algebra I EOC test were compared to the sample items for 2007. The 2007 set of test items were found to be in proportion to the type and number of items on a 2009 Algebra I EOC test. (The thinking skills for the 2009 Algebra I EOC test were not released to the public and could not be used for this research). However, the 2001 sample test
items were not in proportion to the 1998, 2007 or 2009 Algebra I EOC tests. Also, some items appeared in more than one year; for example, several Algebra I test items in 1998 were included in 2001; similarly, several test items in 2001 were included in 2007. However, each set of test items were analyzed independently.

**Results**

The frequency for which the *Dimensions of Thinking* and *Bloom’s taxonomy* thinking skill classifications were used by NC DPI for Algebra I EOC test items varied over the years (see Table 1).

Table 1

*NC DPI’s classification of Algebra I EOC test items for 1998, 2001, and 2007*

<table>
<thead>
<tr>
<th>Thinking Skills assigned by NC DPI to Algebra I test items</th>
<th>1998 # (%)</th>
<th>2001 # (%)</th>
<th>2007 # (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
<td>10 (6%)</td>
<td>2 (3%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Organizing</td>
<td>2 (1%)</td>
<td>4 (6%)</td>
<td>5 (7%)</td>
</tr>
<tr>
<td>Applying</td>
<td>88 (54%)</td>
<td>26 (37%)</td>
<td>40 (56%)</td>
</tr>
<tr>
<td>Analyzing</td>
<td>46 (28%)</td>
<td>24 (34%)</td>
<td>27 (38%)</td>
</tr>
<tr>
<td>Generating</td>
<td>5 (3%)</td>
<td>1 (1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Integrating</td>
<td>9 (6%)</td>
<td>13 (19%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Evaluating</td>
<td>2 (1%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>162 (99%)</strong>*</td>
<td><strong>70 (100%)</strong></td>
<td><strong>72 (101%)</strong>*</td>
</tr>
</tbody>
</table>

* Percent ≠ 100% due to rounding

In 1998, all seven thinking skills were used at least twice; however, by 2007, only three thinking skills (*organizing*, *applying* and *analyzing*) were used. According to an NC DPI official, this was done to make the process of classifying test items more coherent and consistent without losing the integrity of incorporating HOT in the EOC tests (this will be discussed in more detail below).
In the researcher’s analysis of the Algebra I EOC test items, three distinct patterns emerged:

1. It was not unusual for the same (or almost identical) test items to be classified using different thinking skills;

2. A large number of test items did not match their description in Dimensions of Thinking or Bloom’s Taxonomy; and

3. The findings of (1) and (2) were primarily restricted to 1998 and 2001; by 2007, the percent of HOT Algebra I test items increased significantly and appeared to be more consistent with the definition of HOT used in this research.

**Inconsistent identification of thinking skills**

The thinking skills from Dimensions of Thinking and Bloom’s Taxonomy were sometimes inconsistently used by NC DPI to classify test items. Examples of inconsistent identification of a thinking skill are included in Table 2. All test items were multiple choice; however, the choices are not included in the table.

**Table 2**

*Sample test items classified by NC DPI using different thinking skills*

<table>
<thead>
<tr>
<th>NC DPI Classifications</th>
<th>Test Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge &amp; Analyzing</td>
<td>Which is the graph of a line with a slope of 2 and a y-intercept of 5?</td>
</tr>
<tr>
<td>Knowledge &amp; Applying</td>
<td>What is the additive inverse of $-\frac{1}{10}$?</td>
</tr>
<tr>
<td>Organizing &amp; Generating</td>
<td>Which expresses the total surface area (including the top and bottom) of a tower of $c$ cubes each having side length $e$?</td>
</tr>
<tr>
<td>Applying &amp; Analyzing</td>
<td>What is $4x + 5y + 9 = 0$ in slope-intercept form?</td>
</tr>
<tr>
<td>Applying &amp; Analyzing</td>
<td>What is the greatest common factor of $15x^4y^7 - 21x^7y^7 + 6x^3y^2$?</td>
</tr>
<tr>
<td>Applying &amp; Analyzing</td>
<td>Solve: $-4 \leq 3x + 2 \leq 5$</td>
</tr>
<tr>
<td>Applying &amp; Integrating</td>
<td>What is the sales tax on a $15,000 boat if the sales tax rate is 6%?</td>
</tr>
</tbody>
</table>
As can be seen from table 2, it was not unusual for NC DPI to classify the same or similar test items as both LOT (knowledge, organizing, and applying) and HOT (analyzing, integrating, and generating).

**Misclassification of thinking skills**

The second pattern that emerged from the analysis of Algebra test items was the thinking skill assigned to test items often did not match their descriptions in *Dimensions of Thinking* or *Bloom’s Taxonomy*. Examples of misclassification of applying, analyzing, generating, integrating, and evaluating are provided below.

**Applying.** NC DPI officially defined applying as “demonstrating prior knowledge within a new situation. The task is to bring together the appropriate information, generalizations or principles that are required to solve a problem.” (Bold added) This definition was considered by NC DPI to be consistent with application as defined in *Bloom’s Taxonomy*. As noted in *Bloom’s Taxonomy* and in NC DPI’s definition of applying, for a test item to be at the level of application or higher, a “new situation” for the student is required. Bloom emphasized in his original work in 1956 and subsequent discussions on this issue (Bloom et al., 1971; Bloom, Madaus, & Hastings, 1981) that application and higher levels in the taxonomy do not refer to test items where only minor changes are made, but otherwise, the procedure was the same to that practiced in class. As Bloom et al. (1981) stated:

> By ‘new problems and situations’ we mean problems and situations which are likely to be new to the student. These are similar to those which were included in the instruction but have some element of newness or unfamiliarity for the student … It is not a new problem or situation if it is exactly like others solved in class except that new quantities or symbols are used (as in mathematics or physics). (p. 233)
According to Bloom, test items that had already been practiced in class would be labeled *comprehension* or possibly *knowledge*. The following table includes examples of test items considered by the researcher to be mislabeled as *applying* by NC DPI.

Table 3
*Sample test items labeled as “applying” by NC DPI*

| Simplify: $|4 + 2|$ | Factor: $6x^2 - 23x + 10$ |
|------------------|------------------------|
| Simplify: $(3b^2c)(8b^3c^6)$ | Solve: $(x - 20)^2 = 100$ |
| Simplify: $(5x + 2) + (2x + 8)$ | Solve: $-2y - 5 > y + 4$ |
| Solve: $(3x + 6)(2x - 1) = 0$ | Solve: $6x - 12x - 4 = 68$ |
| Simplify: $-2xy(-3xy^2 + 4x^2y)$ | Factor: $6x^2 - 23x + 10$ |

The test items in Table 3 are likely to be routine and familiar to students after taking Algebra I and do not involve a “new situation” or require a student to “solve a problem”; therefore, these items would not be considered *application* in *Bloom’s Taxonomy* nor do they appear to fit NC DPI’s own definition of *applying*.

Although NC DPI began using *Dimensions of Thinking* and *Bloom’s Taxonomy* in the mid-1990s, it is clear that items labeled as *applying* did not fit Bloom’s (1956) definition of *application*. However, NC DPI’s definition of *applying* by 2007 does appear to fit the definition of *apply* based on Anderson, Krathwhol et al.’s (2001) revision of *Bloom’s Taxonomy* where the new version of *apply* is characterized as:

… closely linked with procedural knowledge” and “… when given an exercise (or set of exercises), students typically perform the procedure with little thought. For example, an algebra student confronted with the 50th exercise involving quadratic equations might simply ‘plug in the numbers and turn the crank. (p. 76)
Although this revised version of apply is not consistent with the original Bloom’s Taxonomy, it appears to be consistent with the identification of test items as applying by NC DPI for Algebra I EOC test items and its identification as LOT.

One of the characteristic of HOT in mathematics is its “newness” to the solver or its non-routine nature. However many of the test items labeled as HOT by NC DPI (i.e., analyzing, generating, integrating, and evaluating) were procedural and routine (i.e., LOT). This was mostly prevalent in 1998 and 2001. Examples of test items the researcher considered to be misclassified as HOT are included below. The multiple choices are omitted on most examples.

**Analyzing.** NC DPI defined analyzing as “clarifying existing information by examining parts and relationships; identifying attributes and components; identifying relationships and patterns; identifying errors.” However, the following items were labeled by NC DPI as analyzing (HOT):

- The formula to find the area of a circle is $A = \pi r^2$. What is the area of a circle if the diameter is 16 cm? (Use 3.14 for $\pi$)
- What is the greatest common factor of $-2t^6s^2 - 2r^3s^3t + 16t^2s^3t$?
- Which of the following is an algebraic expression for “twice the sum of a number and 5”?
- What are the coordinates for points R and T respectively?

**Generating.** NC DPI defined generating as “producing new information, meaning, or ideas, going beyond available information to identify what reasonably may be true; anticipating next events, or the outcome of a situation; explaining by adding details, examples, or other relevant information.” However, the following items were labeled by NC DPI as generating (HOT):

- The sum of a number and ten is twelve.” What is the equation for this statement?
• There are 24 yards of rope with which to enclose a rectangular area. If \( w \) is the width of the rectangle, what is the area function for the roped-off rectangle?

**Integrating.** NC DPI defined *integrating* as “combining information efficiently into a cohesive statement and changing existing knowledge structures to incorporate new information.” The following items were labeled by NC DPI as *integrating* (HOT):

• The length of a house is 68 feet and the width is 24 feet. Find the area of the house.

• What is the sales tax on $10,200 automobile if the sales tax rate is 4%?

**Evaluating.** NC DPI defined *evaluating* as “assessing the reasonableness and quality of ideas; creating standards for making judgments; confirming the accuracy of claims.” The following items were labeled by NC DPI as *evaluating* (HOT):

• Which expression is equal to \( \sqrt{144} \)?
  
  a) \( \sqrt{72} + \sqrt{72} \)  
  b) \( \sqrt{100} + \sqrt{4} \)  
  c) \( \sqrt{100} - \sqrt{4} \)  
  d) \( \sqrt{225} - \sqrt{64} \)

• Using different data, two scientists each developed an equation for the same experiment. The equations were \( y = \frac{2}{3}x - 4 \) and \( y = 3x + 10 \). Which ordered pair is valid for both scientists?
  
  a) \(-6, -8\)  
  b) \(\frac{8}{3}, 2\)  
  c) \(-\frac{10}{3}, 0\)  
  d) \(2, 16\)

Similar to *application*, test items labeled as *analyzing, generating, integrating,* and *evaluating* on the Algebra I EOC often did not involve a “new situation” or require students to “solve a problem.” Students taking Algebra I in North Carolina would have seen and practiced tasks the same as or very similar to the items provided above prior to taking the EOC test. As such, this researcher considers these items to be “mislabeled” as HOT. In addition, in some of the examples above, it appears that test items were classified using common mathematical and pedagogical definitions of these terms in place of their meaning in *Dimensions of Thinking* (i.e., *generating* an equation, *integrating* subject matter, or *evaluating* an expression.)
Increase in Algebra I EOC HOT test items over time

As discussed above, numerous Algebra I test items were considered by the researcher to be “mislabeled” as HOT by the NC DPI; however, this was primarily limited to test items in 1998 and 2001; by 2007 only three thinking skills were used: organizing, applying, and analyzing. In these cases, the vast majority of test items labeled analyzing (the only HOT thinking skill used in 2007) appeared to be classified consistently with the definition of HOT used by the researcher. Table 4 compares the NC DPI and researcher classifications between 1998 and 2007.

Table 4

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LOT</td>
<td>61%</td>
<td>86%</td>
<td>46%</td>
<td>76%</td>
<td>63%</td>
<td>64%</td>
</tr>
<tr>
<td>HOT</td>
<td>38%*</td>
<td>14%</td>
<td>54%</td>
<td>24%</td>
<td>38%*</td>
<td>36%</td>
</tr>
</tbody>
</table>

* Percent ≠ 100% due to rounding

In 1998, NC DPI classified 38% of test items as HOT while only 14% were classified as HOT by the researcher. However, by 2007, NC DPI classification of test items as HOT was more consistent with the researcher’s classifications. In addition to reducing the number of thinking skills from seven to three by 2007, the thinking skills of organizing, applying and analyzing were redefined specifically for mathematics. For NC DPI, the thinking skills organizing and applying continued to represent LOT and the thinking skill analyzing continued to represent HOT. Organizing was re-characterized as “performing straightforward computations.” Applying was re-characterized as “solving routine problems;” that is, students do not have to work outside the familiar, but these include tasks that involve more than just computation. Analysis was re-characterized as solving non-routine problems and where they were likely to work outside the familiar (i.e., HOT). Although it appears these descriptions of
organizing, applying, and analyzing fit with the 2007 version of Algebra I EOC, this was not the case for 1998 and 2001.

Re-classifying 1998 / 2001 HOT items as LOT in 2007

The type of test items labeled by NC DPI as either LOT or HOT changed between 1998 and 2007. Two key observations included: (1) the relationship between a test item placed in a real-world context and its categorization as HOT or LOT, and (2) how test items involving translation, representation and interpretation of data were categorized as HOT or LOT.

On the 1998 (Form S) Algebra I EOC test, 20 out of 81 test items were placed in a real-world context, with 17 (85%) labeled HOT by NC DPI (mostly as analyzing or integrating). In 2007, 44 out of 72 test items in were placed in context with 21 (48%) labeled as HOT by NC DPI (i.e., analyzing). (See table 5)

Table 5

| Items placed by NC DPI in a real-world context, 1998 (Form S) vs. 2007 |
|--------------------------|----------|----------------|----------|
|                         | 1998 (Form S)* | 2007       |
|                         | 81 items total | 72 items total |
|                         | (20 in real-world context) | (44 in real-world context) |
| LOT / HOT as identified by NC DPI | LOT | HOT | LOT | HOT |
| No. in real-world context | 3   | 17  | 23  | 21  |

* Only 1998 Form S was used in order to compare a similar number of test items

In 1998, NC DPI appeared to have used a real-world context as a proxy for HOT; this was not the case in 2007 with a slight majority of real-world test items labeled as LOT. The following are example test items from 1998 in a real-world context considered in this research to be mislabeled by NC DPI as HOT:

- The formula to find the area of a circle is \( A = \pi r^2 \). What is the area of a circle if the diameter is 16 cm? (Use 3.14 for \( \pi \)). [Analyzing]
• The length of a house is 68 feet and the width is 24 feet. Find the area of the house. [Integrating]

• What is the sales tax on a $10,200 automobile if the sales tax rate is 4%? [Integrating]

It is a common, but false, assumption that if a task or a test item is placed in a real-world context, then it necessarily engages students in HOT. This assumption appears to be the case for 1998 test items. By 2007, the test items writers appeared to have developed an awareness that a test item having a real-world context does not mean it necessarily engages students in HOT.

Although the percentage of test items labeled as applying were the same in 1998 and 2007 (approximately 55%), many of the test items labeled as applying in 2007 would have been labeled as analyzing, integrating or generating in 1998 and 2001. As an example, the following test item was labeled analyzing (HOT) in 1998 while an almost identical test item was labeled applying (LOT) in 2007

• An object is blasted upward at an initial velocity, \( v_o \), of 240 ft/s. The height, \( h(t) \), of the object is a function of time, \( t \) (in seconds), and is given by the formula \( h(t) = v_o t - 16t^2 \). How long will it take the object to hit the ground after takeoff?

In the 1998 and 2001 Algebra I EOC, many of the test items that focused on the representation and interpretation of data such as matching equations with data in table form, finding slopes of table values, or finding lines (equations) of best fit were labeled as HOT. The two examples below are 2001 Algebra I EOC test items labeled as integrating (HOT):

• In planning a summer trip for his family, Jill’s dad made up a mileage and cost table.

<table>
<thead>
<tr>
<th>Mileage</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 miles</td>
<td>$67.45</td>
</tr>
<tr>
<td>150 miles</td>
<td>$102.45</td>
</tr>
<tr>
<td>175 miles</td>
<td>$111.20</td>
</tr>
</tbody>
</table>

If the relationship between mileage and cost is linear, what equation would apply?

a) \( C = 50m + 50 \)  b) \( C = 0.35m + 49.95 \)  c) \( C = 0.35m - 50 \)  d) \( C = 0.50m + 49.95 \)
• A spring stretches linearly as weight (sic) is added. The table shows data collected for a certain spring.

<table>
<thead>
<tr>
<th>Weight (g)</th>
<th>Stretch (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.5</td>
</tr>
<tr>
<td>500</td>
<td>2.5</td>
</tr>
<tr>
<td>800</td>
<td>4.0</td>
</tr>
<tr>
<td>900</td>
<td>4.5</td>
</tr>
<tr>
<td>1,200</td>
<td>6.0</td>
</tr>
</tbody>
</table>

What is the slope of the line that fits these data for the spring?

a) \( \frac{1}{200} \)  
b) \( \frac{1}{100} \)  
c) \( \frac{1}{50} \)  
d) \( \frac{1}{2} \)

By 2007, similar test items were instead labeled applying (LOT) as opposed to analyzing (or integrating). For example, the following item were labeled as applying by NC DPI in 2007:

• If a linear regression model is fit to the data in the table below, which equation would best represent the data? (Let \( x \) = the number of years after 1960).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Doctors (y)</td>
<td>2,937</td>
<td>3,511</td>
<td>3,754</td>
<td>4,173</td>
<td>4,741</td>
<td>5,019</td>
<td>5,102</td>
</tr>
</tbody>
</table>

a) \( y = 1.01x - 3500 \)  
b) \( y = 82x + 2937 \)  
c) \( y = 83x + 2929 \)  
d) \( y = 83x + 2944 \)

In 2007, to be classified as HOT (i.e., analyzing) for these types of test items, students were required to use or extend the data or understand the parts of equations conceptually. For example:

• The table shows the relationship between calories and fat grams contained in orders of fried chicken from various restaurants.

<table>
<thead>
<tr>
<th>Calories</th>
<th>305</th>
<th>410</th>
<th>320</th>
<th>500</th>
<th>510</th>
<th>440</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat Grams</td>
<td>28</td>
<td>34</td>
<td>28</td>
<td>41</td>
<td>42</td>
<td>38</td>
</tr>
</tbody>
</table>

Assuming the data can best be described by a linear model, how many fat grams would be expected to be contained in a 275-calorie order of fried chicken?

a) 28   b) 27   c) 25   d) 22
• The table below shows the price of rings for various weights of gemstones.

<table>
<thead>
<tr>
<th>Weight (x)</th>
<th>0.17</th>
<th>0.25</th>
<th>0.28</th>
<th>0.35</th>
<th>0.32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price (y)</td>
<td>$355</td>
<td>$642</td>
<td>$823</td>
<td>$1,086</td>
<td>$919</td>
</tr>
</tbody>
</table>

Which statement best interprets the meaning of the y-intercept of the linear function that best fits the data?

a) The price of the ring per unit of weight of the gemstone  
b) The weight of the gemstone per dollar  
c) The cost of the ring with no gemstone  
d) The weight of the gemstone in the ring that costs $0.

Overall, between 1998 / 2001 and 2007, there was a distinct shift in the interpretation of LOT and HOT with many test items involving real-world contexts or requiring students to interpret data in table form being re-labeled from HOT in 1998 / 2001 to LOT by 2007.

**Summary of Results**

Previous research in the 1990s and early 2000s indicated that although most US states created challenging standards, exams developed by states to assess these standards were not very challenging (Lane, 2004; Nichols & Sugrue, 1999; Webb, 1999; 2002; 2007). This research showed that during this same time period, North Carolina had difficulty creating Algebra I EOC test items that assessed for HOT. In addition, this research found that in 1998 and 2001, in NC DPI’s initial effort to use *Dimensions of Thinking* and *Bloom’s Taxonomy* was not consistently applied to similar test items and were often not consistent with how they are defined in *Dimensions of Thinking* and *Bloom’s Taxonomy*. In 1998 and 2001, the majority of Algebra I EOC test items classified as HOT by NC DPI were tasks for which students were very likely to have been taught an algorithm or procedure to solve. By 2007, there was a distinct shift in how test items were classified. In particular, NC DPI reduced the original seven thinking skills to three (*organizing*, *applying*, and *analyzing*) and provided mathematics specific definitions for each of these thinking skills. As a result, there was more consistency in categorizing test items
as well as an increase in the cognitive demands of test items. Overall, compared to the 1998 / 2001 Algebra I EOC test items, the 2007 test items were more complex, did not consider real-world contexts synonymous with HOT, involved more problem solving and conceptual understanding, and required students to solve more test items “outside the familiar.”

Discussion
This case study provides several key findings that other test writers (whether for the classroom or large-scale assessments) can apply to their own situation regarding assessing for HOT. These include the importance of: (a) considering the “familiarity” a student has with a test item in deciding whether the item is LOT or HOT; (b) using a mathematics specific assessment framework with a limited number of categories; and (c) not confusing test items placed in a “real-world” context with HOT.

Role of “familiarity” in defining HOT
Bloom et al. (1956) provides insight into the difficulty experienced by NC DPI in 1998 and 2001 regarding identifying tasks as HOT or LOT. Bloom et al. (1956) stated that when testing for higher levels of thinking, “…the problem situation must be new, unfamiliar, or in some way different from those used in the instruction” (p. 238). In this study, the differences between LOT and HOT are consistent with the differences in Bloom’s Taxonomy between comprehension and application. Comprehension is characterized in Bloom’s Taxonomy as using known algorithms. In contrast, application involves using knowledge in new situations and solving problems without being told which algorithm or method to use. Thus, when creating test items that are likely to engage students in HOT, it is important to take into account the level of familiarity students have with the algorithms, methods of solving the task, or the context of the test item. NC DPI did not appear to do this in 1998 / 2001, and as a result, misinterpreted
Dimensions of Thinking and Bloom’s Taxonomy and over-estimated the amount of HOT on its Algebra I EOC tests. This appears to have been corrected by 2007.

Use mathematics specific frameworks with a limited number of categories

When designing test items, using generic assessment frameworks such as Bloom’s Taxonomy and Dimensions of Thinking can be problematic for two reasons. First, they are not mathematics specific which can result in different and conflicting interpretations of the categories; and second, there are many categories which can result in inconsistent application of the assessment framework. As found in this research, when test item writers interpreted Dimensions of Thinking for mathematics, it appears that test items were often classified using common mathematical and pedagogical definitions of these terms in place of their original meaning in Dimensions of Thinking; for example, generating an equation, integrating subject matter, or evaluating an expression. In particular, research indicates that misapplication of thinking skills in assessment frameworks (e.g., Bloom’s Taxonomy) for mathematics is not uncommon (Gierl, 1997; Thompson, 2008).

Those wishing to create test items that are likely to engage students in HOT should consider using mathematics specific assessment frameworks from national and international exams such as NAEP (NAGB, 2008) and TIMSS (Mullis et al., 2009) or classroom assessment frameworks such as Wilson (1971) and Smith and Stein (1998). These frameworks leave less room for ambiguity than Bloom’s Taxonomy and Dimensions of Thinking. Outlines of these frameworks are shown in figure 1.
ANALYSIS OF HIGHER-ORDER THINKING

Figure 1

*Mathematics assessment frameworks*

<table>
<thead>
<tr>
<th>TIMSS (2011)</th>
<th>NAEP (2009)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Knowing</td>
<td>• Low Complexity</td>
</tr>
<tr>
<td>• Applying</td>
<td>• Moderate Complexity</td>
</tr>
<tr>
<td>• Reasoning</td>
<td>• High Complexity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Computation</td>
<td>• Lower-level demands</td>
</tr>
<tr>
<td>• Comprehension</td>
<td>o Memorization</td>
</tr>
<tr>
<td>• Application</td>
<td>o Procedures without connections</td>
</tr>
<tr>
<td>• Analysis</td>
<td>• Higher-level demands</td>
</tr>
<tr>
<td></td>
<td>o Procedures with connections</td>
</tr>
<tr>
<td></td>
<td>o Doing mathematics</td>
</tr>
</tbody>
</table>

*Additional details are located in the Appendix

In addition, all four of the assessment frameworks shown in figure 1 are not only mathematics specific, but have no more than four categories in which to classify tasks. By 2007, NC DPI reduced the number of categories from seven to three and redefined the categories to focus specifically on mathematics; this resulted in greater consistency in categorizing items as well as increasing the cognitive demands of test items. It is unknown if the assessment frameworks listed above are more effective in teaching and assessing for HOT, but Bloom et al. (1971) noted that Bloom’s Taxonomy was not meant to be used “as is” but should be adapted to meet the needs of individual disciplines. In particular, Wilson’s (1971) assessment framework is an adaptation of Bloom’s Taxonomy. Regardless of the assessment framework one uses, it is also important to have “exemplars” or model tasks or test items that demonstrate each category. All four of the assessment frameworks in figure 1 have companion publications that include sample tasks which can be used to apply the framework consistently and accurately.
HOT and real-world contexts

As noted earlier, it is a common assumption that if a test item is placed in a real-world context, then it must engage students in HOT. However, this is not the case. Consider the following sample test items from NAEP (NAGB, 2008) and Smith (2001).

### NAEP

**Low Complexity, real-world context**

The prices of gasoline in a certain region are $1.41, $1.36, $1.57, and $1.45 per gallon. What is the median price per gallon for gasoline in this region?

A. $1.41  
B. $1.43  
C. $1.44  
D. $1.45  
E. $1.47

**High Complexity, without real-world context**

Which of the following is false for all values of $x$ if $x$ is any real number?

A. $x < x^2 < x^3$  
B. $x^3 < x < x^2$  
C. $x^2 < x < x^3$  
D. $x < x^3 < x^2$  
E. $x^3 < x^2 < x$

### Smith (2001)

**Lower-level demands (procedures without connections), real-world context**

The cost of a sweater at J.C. Penny was $45.00. At the “Day and Night Sale” it was marked 30% off of the original price. What was the price of the sweater during the sale?

**Higher-level demands (doing mathematics), without real-world context**

0.08  0.8  0.08  0.008

Make three observations about the relative size of the four numbers above. Be sure to explain your observations as clearly as possible.

These examples re-enforce the importance of exemplars to help interpret and apply the assessment frameworks accurately and consistently.
Future of NC DPI assessments

In 2010, NC DPI adopted the new Core Curriculum State Standards (CCSS) for Mathematics (http://www.corestandards.org/). For 2012 – 2014, NC DPI will develop its own assessments for these standards; however, beginning in 2015, a nation-wide examination will be given in states who adopted these standards. It will be interesting to see (a) how test items created by NC for CCSS in 2012 – 2014 compare to previous NC DPI EOC tests and (b) how future CCSS examinations compare to NC DPI’s EOC tests.

Final Comments

State, national, and international assessments are important since they serve the role of communicating what teachers should be teaching and what students should be learning (Lane, 2004; McMillan et al., 1999; Nichols & Sugrue, 1999). However, large-scale assessments that fail to live up to their stated claims of cognitive complexity can undermine curriculum reforms at the classroom level. National and international assessments in mathematics indicate that the US is not preparing students to solve complex problems or in general, to think at higher levels (Baldi et al., 2007; Lee et al., 2007; Kloosterman, 2010; Mullis et al., 2008). Although this study focused on only one US state, this research sheds light on potential problems with US state assessments and how this might contribute to low US scores on national and international assessments. While most studies in assessing for HOT look at teachers’ classroom assessment, this research shows that many of the same difficulties classroom teachers face can occur on larger-scale assessments. Thus, more attention should be paid to state, national, and international exams and whether these tests assess what they claim to assess.
Limitations

These findings are restricted to one US state’s effort to develop items for its Algebra I EOC tests. Therefore, no claim is made for current North Carolina EOC tests or other US state exams. In addition, other researchers may interpret LOT or HOT differently and therefore, arrive at different conclusions regarding the test items analyzed in this study. It is also important to emphasize that the purpose of this research was not to evaluate NC’s current testing program, but to better understand, from an historical perspective, how one US state used *Dimensions of Thinking* and *Bloom’s Taxonomy* to assess HOT for Algebra I EOC test items. As such, this research contributes to the current knowledge base on the challenges of using assessment frameworks (e.g., *Dimensions of Thinking* and *Bloom’s Taxonomy*) to assess for HOT in mathematics.
References


APPENDIX

MATHEMATICS ASSESSMENT FRAMEWORKS

TIMSS 2011 ASSESSMENT FRAMEWORK (Mullis et al., 2009)

Knowing

(a) Recall: Recall definitions; terminology; number properties; geometric properties; and notation (e.g., \( a \times b = ab, a + a + a = 3a \)).

(b) Recognize: Recognize mathematical objects (e.g., shapes, numbers, expressions, and quantities). Recognize mathematical entities that are mathematically equivalent (e.g., equivalent familiar fractions, decimals and percents; different orientations of simple geometric figures).

(c) Compute: Carry out algorithmic procedures for \(+, -, \times, \div\), or a combination of these with whole numbers, fractions, decimals and integers. Approximate numbers to estimate computations. Carry out routine algebraic procedures.

(d) Retrieve: Retrieve information from graphs, tables, or other sources; read simple scales.

(e) Measure: Use measuring instruments; choose appropriate units of measurement.

(f) Classify / Order: Classify / group objects, shapes, numbers, and expressions according to common properties; make correct decisions about class membership; order numbers and objects by attributes.

Applying

(a) Select: Select an efficient / appropriate operation, method, or strategy for solving problems where there is a known procedure, algorithm, or method of solution.

(b) Represent: Display mathematical information and data in diagrams, tables, charts, or graphs, and generate equivalent representations for a given mathematical entity or relationship.

(c) Model: Generate an appropriate model (e.g., equation, geometric figure, or diagram) for solving a routine problem.

(d) Implement: Implement a set of mathematical instructions (e.g., draw shapes and diagrams to given specifications).

(e) Solve Routine Problems: Solve standard problems similar to those encountered in class. The problems can be in familiar contexts or purely mathematical.

Reasoning

(a) Analyze: Determine, describe, or use relationships between variables or objects in mathematical situations, and make valid inferences from given information.

(b) Generalize / Specialize: Extend the domain to which the result of mathematical thinking and problem solving is applicable by restating results in more general and more widely applicable terms.

(c) Integrate / Synthesize: Make connections between different elements of knowledge and related representations, and make linkages between related mathematical ideas. Combine
mathematical facts, concepts, and procedures to establish results, and combine results to produce a further result.

(d) Justify: Provide a justification by reference to known mathematical results or properties.

(e) Solve Non-routine Problems: Solve problems set in mathematical or real life contexts where students are unlikely to have encountered closely similar items, and apply mathematical facts, concepts, and procedures in unfamiliar or complex contexts.

NAEP 2009 ASSESSMENT FRAMEWORK (NAGB, 2008)

Low complexity

- Students to recall or recognize concepts or procedures.
- Items typically specify what the student is to do, which is often to carry out some procedure that can be performed mechanically.
- The student does not need to use an original method or to demonstrate a line of reasoning.

Moderate complexity

- Items involve more flexibility of thinking and choice among alternatives than do those in the low complexity category.
- Student is expected to decide what to do and how to do it, bringing together concepts and processes from various domains. For example, student may need to represent a situation in more than one way, to draw a geometric figure that satisfies multiple conditions, or to solve a problem involving multiple unspecified operations.
- Students might be asked to show explain their work but would not be expected to justify it mathematically.

High complexity

- Students are expected to use reasoning, planning, analysis, judgment, and creative thought.
- Students may be expected to justify mathematical statements or construct a mathematical argument. Items might require students to generalize from specific examples.
- Items at this level take more time than those at other levels due to the demands of the task, not due to the number of parts or steps.
WILSON (1971) ASSESSMENT FRAMEWORK

Computation
- Knowledge of specific facts
- Knowledge of terminology
- Ability to carry out algorithms

Comprehension
- Knowledge of concepts
- Knowledge of principles, rules, and generalizations
- Knowledge of mathematical structure
- Ability to transform problem elements from one mode to another
- Ability to follow a line of reasoning
- Ability to read and interpret a problem

Application
- Ability to solve routine problems
- Ability to make comparisons
- Ability to analyze data
- Ability to recognize patterns, isomorphisms, and symmetries

Analysis
- Ability to solve non-routine problems
- Ability to discover relationships
- Ability to construct proofs
- Ability to criticize proofs
- Ability to formulate and validate generalizations
SMITH & STEIN (1998) ASSESSMENT FRAMEWORK

Lower-level demands (memorization)

- Involve either reproducing previously learned facts, rules, formulas, or definitions or committing facts, rules, formulas, or definitions to memory.
- Cannot be solved using procedures because a procedure does not exist or because the time frame in which the task is being completed is too short to use a procedure.
- Are not ambiguous. Such tasks involve the exact reproduction of previously seen materials, and what is to be produced is clearly and directly stated.
- Have no connection to the concepts or meaning that underlie the facts, rules, formulas, or definitions being learned or reproduced.

Lower-level demands (procedures without connections)

- Are algorithmic. Use of the procedure either is specifically called for or is evident from prior instruction, experience, or placement of the task.
- Require limited cognitive demand for successful completion. Little ambiguity exists about what needs to be done and how to do it.
- Have no connection to the concepts or meaning that underlie the procedure being used.
- Are focused on producing correct answers instead of on developing mathematical understanding.
- Require no explanations or explanations that focus solely on describing the procedure that was used.

Higher-level demands (procedures with connections)

- Focus students' attention on the use of procedures for the purpose of developing deeper levels of understanding of mathematical concepts and ideas.
- Suggest explicitly or implicitly pathways to follow that are broad general procedures that have close connections to underlying conceptual ideas as opposed to narrow algorithms that are opaque with respect to underlying concepts.
- Usually are represented in multiple ways, such as visual diagrams, manipulatives, symbols, and problem situations. Making connections among multiple representations helps develop meaning.
- Require some degree of cognitive effort. Although general procedures may be followed, they cannot be followed mindlessly. Students need to engage with conceptual ideas that underlie the procedures to complete the task successfully and that develop understanding.

Higher-level demands (doing mathematics)

- Require complex and non-algorithmic thinking – a predictable, well-rehearsed approach or pathway is not explicitly suggested by the task, task instructions, or a worked-out example.
• Require students to explore and understand the nature of mathematical concepts, processes, or relationships.

• Demand self-monitoring or self-regulation of one's own cognitive processes.

• Require students to access relevant knowledge and experiences and make appropriate use of them in working through the task.

• Require students to analyze the tasks and actively examine task constraints that may limit possible solution strategies and solutions.

• Require considerable cognitive efforts and may involve some level of anxiety for the student because of the unpredictable nature of the solution process required.