



CALA Report 108

The Capabilities-Complexity Model

Albert Oosterhof
Faranak Rohani
Carol Sanfilippo
Peggy Stillwell
Karen Hawkins

June 2008



Center for Advancement of Learning and Assessment
Florida State University, Tallahassee, FL
www.cala.fsu.edu

THE CAPABILITIES-COMPLEXITY MODEL

The Capabilities-Complexity Model

Albert Oosterhof, Faranak Rohani, Carol Sanfilippo,
Peggy Stillwell, and Karen Hawkins

Center for Advancement of Learning and Assessment
Florida State University

Acknowledgments

Center for Advancement of Learning and Assessment (CALA)—As a pioneer in research and the design of multimedia instructional materials and customized assessments, CALA assists policy makers and educators by providing them analyses and practical solutions.

The research described herein was funded by the Florida Department of Education. The findings and opinions do not necessarily reflect the positions or policies of the Florida Assessment and School Performance Office or the Florida Department of Education.

Editor: Alice Fisher

Media Specialist: Liane Schuessler

Copyright © 2008 by the Center for Advancement of Learning and Assessment, Florida State University. All rights reserved.

The Capabilities-Complexity Model

In assessment, the ability to construct test items that measure a targeted skill is fundamental to validity and alignment. The ability to do the reverse is also important: determining what skill an existing test item measures. This paper proposes a model for classifying test items that builds on procedures developed by others, including Bloom (1956) and Webb (2002). An advantage of the proposed model is that it references both the type of cognitive ability involved and the complexity of skill being assessed. This model is referred to as the Capabilities-Complexity Model (CCM).

Implementing the CCM involves two sequential steps. First, the type of capability involved with each test item is established. Capability refers to the mental process a person uses to complete a particular cognitive task, such as answering a test item. This process, though essential and very real to the person completing it, cannot be observed directly by another person. The CCM addresses three types of capabilities: declarative knowledge, procedural knowledge, and problem solving. Identifying the type of capability is relevant because different types of performance are used to assess proficiency with the different capabilities.

After the type of capability is identified for each item, the level of cognitive complexity is established. The CCM divides complexity into three levels. To establish a test item's complexity, different qualities are examined for items measuring declarative knowledge, procedural knowledge, and problem solving. This is why the type of capability is determined before establishing the level of complexity. Determining the complexity of items helps ensure that a test includes more cognitively demanding items in addition to those of simpler complexity. Although the CCM involves both capability and complexity dimensions, using the model may actually simplify the classification of test items because examining an item's complexity becomes more focused once the type of capability is identified.

This paper first provides a historical perspective for the CCM. It then describes how the CCM is used to classify test items and provides initial findings related to using the model.

Background

Probably the most widely used approach for classifying test items has been the taxonomy of educational objectives proposed by Bloom (1956). Bloom's six categories—knowledge, comprehension, application, analysis, synthesis, and evaluation—are referenced in hundreds of professional papers, as well as in many books concerned with student assessment. Categories from Bloom's taxonomy are often used to develop test plans.

Less widely known is that Bloom subdivided the taxonomy into two primary categories. The first primary category is knowledge; it is also the first of his six familiar classifications. Bloom's knowledge category corresponds closely to what contemporary cognitive psychologists refer to as declarative knowledge. The second primary category is what Bloom originally called "intellectual abilities and skills" and later simply "intellectual skills." This category is represented by the subcategories of comprehension, application, analysis, synthesis, and evaluation. In many ways, Bloom's intellectual skills category approximates what cognitive psychologists refer to as procedural knowledge, although his five subcategories depart substantially from what is now known about procedural knowledge (Confrey, 1990; Gierl, 1997; Snow, 1989; Tittle, Hecht, & Moore, 1993).

The discrepancies between Bloom’s intellectual skills subcategories and those used more recently in cognitive psychology might explain some of the problems often experienced when using the taxonomy to classify test items in large-scale assessments. Educational agencies sometimes provide vague explanations for these problems, such as suggesting that Bloom’s taxonomy involves assumptions about “the student’s approach to the item” in contrast to the more determinable “cognitive demand inherent in the test items”¹ and the need to “focus on the expectations of the item, not the ability of the student.”²

A useful alternative for classifying test items is Webb’s (2002) Depth of Knowledge (DOK) model. Webb identifies four levels of complexity, varying their descriptions across content areas. For example, the levels of complexity in science are

- *Level 1—Recall and Reproduction*, such as recalling facts and terms, using words to represent scientific concepts and relationships, and performing a routine procedure such as measuring length;
- *Level 2—Skills and Concepts*, such as explaining relationships between facts and variables, describing and explaining examples of a scientific concept, and organizing and interpreting data;
- *Level 3—Strategic Thinking*, such as identifying a research question and designing an investigation for a scientific problem, solving a nonroutine problem, and forming conclusions from experimental data; and
- *Level 4—Extended Thinking*, which involves complex reasoning, relating ideas within or between content areas, and devising a way to solve a complex problem in a situation that could involve a variety of potential strategies. Webb points out that completing a Level 4 task would likely require an extended period and multiple opportunities for observation. His example of a Level 4 task involves conducting a scientific investigation that includes specifying a problem, designing and carrying out an experiment, analyzing the data, and forming conclusions.

Unlike Webb’s classification, Bloom’s (1956) taxonomy does not explicitly reference levels of complexity. However, varying levels of cognitive complexity are implicit in Bloom’s illustrations, such as the category of knowledge. Many educators mistakenly believe that knowledge refers to limited facts and similar knowledge that can be learned through memorization. Bloom does include factual knowledge as a type of knowledge, but he also includes these additional categories of knowledge:

- “ways of organizing, studying, judging, and criticizing ideas and phenomena” (Bloom, 1956, p. 68);
- “criteria by which facts, principles, opinions, and conduct are tested or judged” (Bloom, 1956, p. 72);
- methodology such as “methods of inquiry, techniques, and procedures employed in a particular subject field” (Bloom, 1956, p. 73);
- “universals and abstractions in an academic or professional field, such as knowledge of particular abstractions which summarize observations of phenomena” (Bloom, 1956, p. 75); and
- theories and structures such as knowledge of “the body of principles and generalizations together with their interrelations which present a clear, rounded, and systematic view of a complex phenomenon, problem, or field” (Bloom, 1956, p. 76).

Bloom's knowledge category incorporates elements of Webb's Level 1 and Level 2 complexity, and possibly some elements of Level 3. A similar range of cognitive complexity is evident in the other categories of Bloom's taxonomy.

As noted earlier, Bloom's taxonomy does not fit well with what is presently understood about the nature of learned knowledge. Gagné, Yekovich, and Yekovich (1993) present a highly useful and readable discussion of cognitive psychology as it relates to school learning. They divide knowledge and skills into three main categories: declarative knowledge, procedural knowledge, and problem solving, which will be discussed in detail later. Presently, note that declarative knowledge involves information that can be declared orally or in writing. Declarative knowledge is roughly equivalent to what Bloom refers to as knowledge.

Procedural knowledge is knowing how to do something, which is quite different from discussing or explaining what is known. For instance, a physics student being able to explain Ohm's Law (the relationship between electrical resistance, current, and voltage) is an example of declarative knowledge, whereas being able to use Ohm's Law to establish whether the resistance of an electrical wire will accommodate particular current and voltage is a demonstration of procedural knowledge. The nature of procedural knowledge is complex. Oosterhof (in press) structures procedural knowledge into making discriminations, understanding concepts, and applying rules that govern relationships. Cognitive psychologists find that procedural knowledge often includes motor skills and cognitive strategies.

Problem solving builds on declarative and procedural knowledge. Problem solving may involve domain-specific strategies, suggesting that different strategies are employed when solving problems in different content areas, such as writing and science. Gagné, Yekovich, and Yekovich (1993) indicate that problem solving exists when one has a goal but has not yet identified a means for reaching it. For example, problem solving would involve the process of writing an expository paper, preparing a persuasive speech, or conducting a scientific investigation. As with Bloom's categories, declarative knowledge, procedural knowledge, and problem solving each varies in terms of cognitive complexity. Likewise, the lower as well as higher levels of cognitive complexity each include a variety of capabilities, with different types of assessment tasks particularly effective at measuring these different types of capabilities. This paper recommends that types of capabilities and levels of complexity should not be controlled by chance when classifying and evaluating the adequacy of a set of test items, or when aligning assessments with standards. This paper also proposes that the task of classifying test items and the development of standards and alignment of tests is simplified if both capabilities and complexities are considered.

Classifying Items with the Capabilities-Complexity Model

Implementing the CCM³ involves two sequential steps:

- First, identify the type of capability involved with each test item.
- Then, establish the level of cognitive complexity for each test item.

Fairly different qualities are considered when examining test items that measure the different capability types. This is why the type of capability involved is first identified for the items, before establishing the level of complexity.

Types of Capabilities

Knowledge consists of different types of capabilities, and different types of tasks or performances are particularly effective at assessing the capability types. When developing or revising assessments, it is important to be aware of the different capabilities, be familiar with the tasks that effectively measure each type, and be able to classify test items according to the type of capability each measures. This paper divides capabilities into three types referenced earlier: declarative knowledge, procedural knowledge, and problem solving.

Nature of Declarative Knowledge. Declarative knowledge is the ability to express information and ideas orally or in writing. These questions involve declarative knowledge:

What is the capital of New York?

In what year did a person first land on the moon?

Although some suggest declarative knowledge is limited to memorizing facts, it includes much more. For instance, the following questions concern declarative knowledge:

What is the difference between libel and slander?

In what ways are a CD-ROM and a DVD-ROM similar? In what ways are they different?

Declarative knowledge includes more complex information such as that associated with the following:

- properties (Explain why much less energy is required to change water from room temperature to near boiling than is required to change its temperature slightly to convert it to steam.)
- principles (Why does the price of a product increase when its demand increases?)
- trends (As the average age of the U.S. population increases, what trends are occurring in health care?)
- terminologies (In economics theory, what is inflation?)
- techniques (Describe how to distill water.)
- phenomena (When the moon eclipses the sun, why can the eclipse be observed on only a small portion of Earth's surface?)

All of these examples deal with declarative knowledge because each question or command is answered by expressing information. Declarative knowledge involves any knowledge that can be discussed or demonstrated through declaration.

Nature of Procedural Knowledge. Procedural knowledge is the ability to perform actions. Unlike declarative knowledge, which involves expressing information, procedural knowledge involves *doing* something. More specifically, it requires applying a principle, relationship, or idea to a new but comparable setting. If a particular procedure has been learned, it can be applied to additional situations for which the procedure is relevant. These are examples of procedural knowledge:

- if one knows how to use flowcharts, interpreting information from a previously unseen flowchart;
- if one knows how to convert temperatures from Fahrenheit to Celsius scales, calculating the Celsius equivalents of Fahrenheit temperatures;

- if one knows the distinguishing characteristics that identify an animal as a bird, identifying which among previously unseen examples of flying and nonflying animals are birds; and
- if one knows what condition will result in an object floating in water, applying that principle to predict whether an unfamiliar object will float if placed in water.

Although both declarative and procedural knowledge may have the same content, declarative knowledge involves stating, describing, discussing, or explaining something, whereas procedural knowledge requires using a learned procedure to do something. Some contrasting examples are

Declarative Knowledge	Procedural Knowledge
In the context of marketing, being able to explain what <i>celebrity promotion</i> means	When shown previously unseen examples of advertising, recognizing which examples illustrate <i>celebrity promotion</i>
Being able to describe how distillation purifies water	Being able to produce purified water using distillation
Being able to explain what is meant by the <i>main idea</i> of a written passage	Being able to read a written passage and identify its main idea

Concepts and *rules* are two important subcategories of procedural knowledge. Procedural knowledge of concepts concerns being able to determine whether something constitutes an example of the concept. For example, a person who has learned the concept of a bird should be able to look at previously unseen photographs of animals such as the eagle, bat, butterfly, sparrow, ostrich, and bee, and indicate which are birds. Concepts can involve abstractions. For instance, a person who knows the concept of anxious should be able to identify examples that illustrate a person who is anxious. Note that procedural knowledge of a concept does not necessitate explaining the concept—that would be declarative knowledge. Instead, the person is invoking a procedure: classifying new illustrations as being examples or nonexamples of the concept.

In contrast to concepts, procedural knowledge of rules involves applying a principle that expresses the relationships among a class of objects or events. Some examples of procedural knowledge pertaining to rules are

- locating a position on a globe when provided its longitudinal and latitudinal values,
- using a formula to calculate the dollar amount of interest on a bank loan, and
- changing verb tense of a sentence from present to past perfect.

Nature of Problem Solving. Problem solving involves achieving a goal when the means for reaching that goal has not yet been established. With problem solving, a person must have a sense of the goal. The person might have been told what the goal is or may have to establish the goal through inference. Reaching the goal—that is, solving the problem—involves applying strategies learned previously as declarative and procedural knowledge.

Problem solving comprises three basic steps:

1. establishing a representation of the problem (that is, establishing a sense of the goal);

2. based on previously learned knowledge, selecting a strategy that seems appropriate for reaching the goal; and
3. employing the strategy and evaluating the results.

The person continues to select and employ strategies until the problem is solved (or until the effort is abandoned). This is an example of problem solving:

1. Represent the problem:
You need a replacement computer for word processing and for accessing the Internet and can spend up to \$600.
2. Select a strategy:
You know your friend has helped others select good, inexpensive computers, so you will seek your friend's advice.
3. Employ the strategy and evaluate results:
You try to contact your friend, but discover the friend is unavailable for three weeks. This is too long a wait, so you must use a different strategy.

In this example, problem solving would continue with alternate strategies until a computer is located or the effort is abandoned.

Note that we are not talking about problem solving in a mathematical sense. Math problems that require multiplication, applying formulas, and even more complicated procedures such as trigonometry and calculus often involve applications of rules, not problem solving in the sense used here. Once a particular mathematical technique is learned, that procedural knowledge can be applied to other math problems that use the procedure. Some problem-solving situations require mathematical procedures. As the terminology is used here, those situations represent problem solving as long as the process involves the sequence of three basic steps listed earlier.

Levels of Complexity

Test items that require recalling facts that can be memorized are less cognitively complex than questions that require an explanation of some phenomenon. Likewise, questions that can be answered by doing one thing are less cognitively complex than questions that require doing many things, particularly if these different things must be addressed simultaneously.

Some test questions are less complex because they are involved with things that can be seen or touched rather than abstractions. For instance, test questions concerned with physical attributes, such as describing how chairs and tables are similar or different, usually are less complex than questions concerned with abstractions, such as describing how patience and kindness are similar or different.

An interesting characteristic of cognitive complexity is that it has little to do with the difficulty of test items. Some items have low complexity because they involve a single action and no abstraction, but are very difficult to answer correctly because they deal with obscure information (e.g., naming the 10 longest rivers in the United States). On the other hand, some cognitively complex tasks are quite easy to complete. For example, finding and ordering a book from a major online retailer is easy to do, even though the process requires a number of steps as well as abstractions. Major online retailers would not be profitable if the ordering process was difficult to complete.

With the CCM, all test items being reviewed are first grouped into the three capability types:

- declarative knowledge
- procedural knowledge
- problem solving

After this grouping, the complexity level is established for the items within each group, usually beginning with the items measuring declarative knowledge. Different attributes are considered when establishing the complexity of declarative, procedural, and problem-solving items. In each case, the items are sorted by complexity into Level 1, Level 2, or Level 3.

Establishing the Complexity Level of Items that Measure Declarative Knowledge.

Declarative knowledge is the ability to convey information and ideas using oral or written expressions. Examples of declarative knowledge include

- recollection of facts such as names of people and dates of events;
- descriptions of characteristics such as types of clothes worn to an event last year or weather conditions at that event;
- trends such as population growth and corresponding changes in health care; and
- explanations such as techniques to perform distillation and phenomena such as ocean tides.

Level of complexity for declarative knowledge is influenced by these four attributes:

<i>Type of action involved</i>	Actions involving recall are less complex than actions involving explanations.
<i>Type of information involved</i>	Information about facts and characteristics is less complex than information involving phenomena and principles.
<i>Concreteness of information involved</i>	Concrete information is less complex than abstract information.
<i>Similarity to the context in which the information was learned</i>	Using information in a setting similar to the one in which it was learned is less complex than when the context is different.

The CCM uses the following rubric to classify declarative knowledge items into the three levels of complexity. All the rubrics in this paper are used holistically. That is, participants select the level with characteristics closest to the complexity of the test item they are trying to classify.

LEVEL	What type of action is involved?	What type of information is involved?	How abstract is the information?	How similar is the context to the one in which the information was learned?
1	Recalling Describing	Facts Characteristics Terminologies Properties Phenomena	Typically concrete	The context is the same as during instruction.
2	Describing Explaining	Properties Phenomena Concepts Principles Techniques	Often abstract	The context is similar or parallel to instruction.
3	Explaining Analyzing Differentiating Synthesizing	Properties Phenomena Concepts Principles Techniques	Abstract	The context is clearly different from instruction.

Here are illustrations of each level of complexity for declarative knowledge:

- Example 1:** Most of the craters on the surface of the moon were formed by
- A. eruptions of volcanoes.
 - B. extinct lakes.
 - C. impacts of asteroids.
 - D. solar flares.

This item asks the test taker to recall a characteristic of the moon. This characteristic is a concrete feature of the moon. The complexity of this item is **Level 1**.

Example 2: The measuring cup shown here will be filled half-full with water, and then the irregularly shaped rock will be placed in the cup. Tell how this allows one to measure the volume of this rock.



As used here, the word *tell* is equivalent to *explain*. The type of information involves a *principle* related to the rock displacing water equivalent to the volume of the rock. This illustration is probably similar, but not identical, to the illustrations used during instruction; that is, the *context is similar or parallel to instruction*. The complexity of this item is **Level 2**.

Example 3: The NASA Mars Exploration Rovers *Spirit* and *Opportunity* were programmed with artificial intelligence that evaluates commands radioed from Earth. This helps ensure that commands sent from Earth do not endanger the robots' mission on Mars. Describe an application on Earth where it would be equally important to provide a robot with artificial intelligence. Explain why it would be important to use artificial intelligence in that application.

This item requires the test taker to *synthesize properties* of artificial intelligence, *explain* the principles behind the technology, and *analyze* its application to another setting. Most likely, the *context is clearly different from that of instruction*. The phenomenon of artificial intelligence is *abstract*. The complexity of this item is **Level 3**.

It is helpful to remember that the complexity of items is established after the type of capability is determined. Also, the level of complexity is established for only items from one capability category. That is, the level of complexity is determined for items that measure declarative knowledge, then items that measure procedural knowledge, and then problem solving. The classification process used by the model may seem somewhat overwhelming. However, the content experts who classify the items initially determine only the type of capability being measured by the respective items. Only then do they need to know how to classify an item's complexity. They must first focus on the items judged to measure declarative knowledge. When that task is complete, they learn how to judge the complexity of items that measure procedural knowledge.

As just illustrated, instruction for using the CCM is always accompanied by practice examples. Initially, examples are used to reinforce what has just been learned—for instance, the difference between complexity levels of items that measure declarative knowledge. Then, additional example test items are provided for which educators are asked to talk through how they established the complexity level of these items. Experience to date suggests that participants can quickly learn the process of using the CCM and correctly apply this process to example test items that are known to represent a particular type of capability or level of complexity.

Establishing the Complexity Level of Items that Measure Procedural Knowledge. Procedural knowledge is different from declarative knowledge. While declarative knowledge involves recalling and explaining things, procedural knowledge involves using learned rules or concepts to do something. Here are examples of procedural knowledge:

- using mathematical relationships such as calculating the area within a circle when given its diameter;
- using a rule such as adding “ed” to English verbs to change tense from present to past;
- using techniques to retrieve information, such as using a library card catalog to locate a book; and
- using knowledge of a concept to classify illustrations, such as establishing whether or not a previously unseen animal is a mammal.

The following five attributes make some procedural knowledge more complex than others:

<i>Number of steps involved in the procedure</i>	Procedures involving a single operation are less complex than those involving multiple steps or multiple simultaneous operations.
<i>Directness of instructions</i>	Procedures are less complex if the tasks to be performed can be specifically stated in instructions rather than requiring inference.
<i>Abstractness of illustrations or variables</i>	Procedures are less complex when they involve tangible rather than abstract illustrations or variables.
<i>How narrowly the procedure guides actions taken with the variables</i>	Procedures that employ a highly specific action are less complex than those for which variations of the action can be employed.
<i>Similarity to the context in which the procedure was learned</i>	Using a procedure in a setting similar to the one in which it was learned is less complex than when the context is different.

The following rubric is used to classify procedural knowledge items into the three levels of complexity. Select the level with the characteristics closest to the complexity of the test item you are trying to classify.

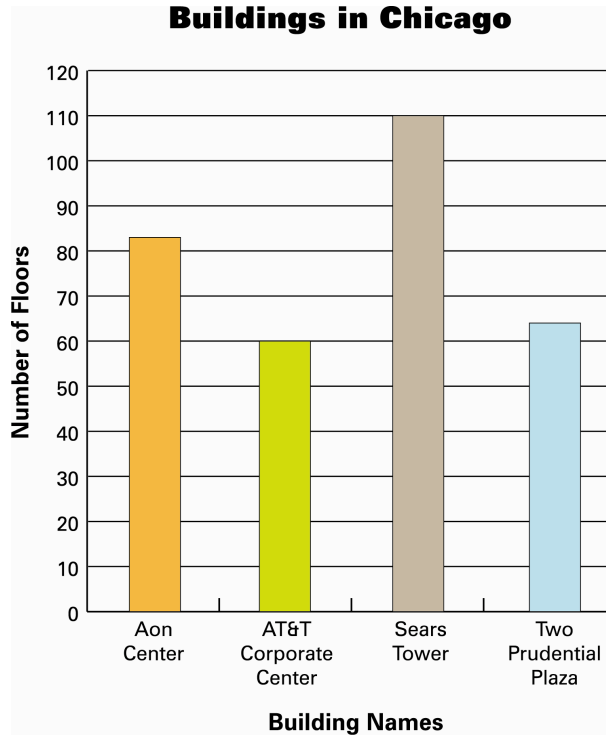
LEVEL	How many steps or operations are involved?	How direct are the instructions on completing the procedure?	How abstract are the illustrations or variables?	How narrowly does the procedure guide actions taken with the variables?	How similar is the context to the one in which the procedure was learned?
1	A single operation is involved.	Specific tasks to be performed are stated directly.	Tangible	There is one specific way to implement the procedure.	The context is likely the same as the one used in instruction.
2	Multiple steps are performed sequentially.	Specific tasks to be performed require some inference.	May be abstract	There is some variation in how to implement the procedure.	The context is likely similar or parallel to instruction.
3	Multiple steps are performed sequentially or simultaneously.	Specific tasks to be performed must be inferred from instructions.	Abstract	There are significant variations in ways to implement the procedure.	The context is likely quite different from instruction.

These test items illustrate each level of complexity for procedural knowledge:

Example 4: This bar graph shows the number of floors in four tall buildings in Chicago.

How many floors does the Aon Center have?

- A. 60
- B. 64
- C. 83
- D. 110



This item represents **Level 1** complexity. Using the graph to find the number of floors of the Aon Center requires a *single operation*, that of projecting the top of the bar in the chart to the number scale on the left. The graph is *tangible*, and the *task to be performed is stated directly*. There is only *one way* to use the top of the bar to establish the number of floors in the Aon Center. With the skill so narrowly focused, the *context presented for interpreting this bar graph is probably the same* as the one used when the skill was learned.

Example 5: The only times a commuter rides the subway are to and from work and to visit a friend. The commuter spends a total of 1.5 hours on the subway each day she commutes to work and an additional 2 hours whenever she visits her friend. Which expression represents her time on the subway if she commutes to work n days and visits her friend 4 times?

- A. $1.5n + (4 \times 2)$
- B. $(1.5 + 2) \times (4 + n)$
- C. $4n(1.5 + 2)$
- D. $4(1.5n + 2)$

According to the rubric, this item is closest to **Level 2** complexity. *Multiple sequential steps* are involved. The variables are *abstract*, and the *directions require some inference*. *Some variation exists in how different people would identify* the correct mathematical expression. There also is some *flexibility in how people might be asked to identify a correct math expression*, although the context would probably parallel the one in which the procedure of using mathematical expressions is taught.

Example 6: Purchase a book from a major online retailer.

This test item is **Level 3** complexity. At first glance, this procedure might seem less complex than the previous example, particularly if you purchase books online often and find the task easy to perform. Be careful, however, not to confuse familiarity with complexity. Using the rubric, you can recognize that purchasing a book online *involves multiple sequential or simultaneous steps with abstract variables* such as titles, shipping information, and quantities within a virtual shopping cart. Also, the *context provided by the Web sites of different retailers can vary considerably and be quite different from the one originally used* for online ordering. The instructions require inference.

Establishing the Complexity Level of Items that Measure Problem Solving. Problem solving focuses on achieving a goal when the means for reaching the goal has not yet been established. Problem solving always involves the sequence of steps described earlier:

1. establishing a representation of the problem (that is, establishing a sense of the goal);
2. based on previously learned knowledge, selecting a strategy that seems appropriate for reaching the goal; and
3. employing the strategy and evaluating the results.

As with declarative and procedural knowledge, some problem-solving tasks are more cognitively complex than others. Four attributes affect this complexity:

<i>Conciseness with which details of the goal to be achieved are stated</i>	Problem solving is less complex when specific details of the goal are provided and little or no inference is needed to establish the problem to be solved.
<i>Variation of strategies typically used to solve the problem</i>	Problems for which a dominant strategy typically is used are less complex than when people use diverse strategies to solve the problem.
<i>Number of steps or operations used to solve the problem</i>	Problems that can be solved in one step or operation are less complex than those that involve multiple steps or operations.
<i>Amount of originality required to solve the problem</i>	Problem solving is less complex when the context in which the strategies are employed is similar to the context in which the strategies were learned.

The following rubric is used to classify problem-solving items into the three levels of complexity. Select the level with characteristics closest to the complexity of the test item you are trying to classify.

LEVEL	How concisely stated are details of the goal to be achieved?	Will a dominant or diverse strategies be used to solve the problem?	How many steps or operations do reasonable strategies involve?	How much originality does solving the problem require?
1	Specific details of the goal to be achieved are stated.	Most people will use the same strategy.	The dominant strategy involves a single step or operation.	Problem-solving strategy will be employed in a context similar to that in which the strategy was used previously.
2	Some details of the goal to be achieved are stated; inference is required to establish the problem to be solved.	Different people will use a limited combination of strategies to solve the problem.	Strategies involve multiple sequential steps or operations.	Problem-solving strategies will be employed in a context somewhat different from that in which they were used previously.
3	The goal to be achieved is stated broadly; inference is required to establish the problem to be solved.	Different people will use numerous and various combinations of strategies to solve the problem.	Strategies involve multiple sequential and/or simultaneous steps or operations.	Problem-solving strategies will be employed in a context significantly different from that in which they were used previously.

Examples 7–9 illustrate the three levels of complexity.

Example 7: Living in Florida, you are surprised and delighted to wake up to a light snowfall. However, you want to immediately remove the snow from the porch and the steps leading to your house to ensure the safety of an approaching visitor. Because you live in Florida, you have nothing designed to remove the snow. What do you do?

This item is **Level 1** complexity. *Specific details of the goal to be achieved are stated*, requiring little inference: there is snow on the porch and steps that must be removed quickly so the visitor will not slip and fall. Therefore, something presently available must be used to remove the snow. *Most everyone will use the same strategy*: sweep the snow off with something like a broom. Using a broom *requires little originality* because a broom likely has been previously used as a strategy for sweeping the porch and steps. Sweeping the porch can be seen as a *single operation*.

Example 8: You are interested in a teaching position that will be available next summer in Poland. You have sent a letter indicating your interest, and in response you have been asked to submit your résumé. Create an updated résumé that will support your application for this position.

This item is **Level 2** complexity. It is clear that a résumé has been requested; however, a reasonable inference would be that the current résumé should be updated to address questions of particular interest to the prospective employer in Poland. That is, *some details of the goal to be achieved are stated, requiring inference to establish the problem to be solved. More than one strategy might be used* to determine what to include in the résumé, such as talking with colleagues who might be familiar with qualifications that will be relevant to the position or using the Internet to find information about the school. The strategy employed for updating the résumé likely *involves multiple operations that will be completed sequentially*, such as seeking and evaluating possible sources of information relevant to the position. The problem of providing an appropriate résumé requires problem-solving strategies used *in a context somewhat different from that in which the strategies were used previously*.

Example 9: You accepted a new administrative position in a school district where you will coordinate curriculum development and evaluation in your content area. Your staff currently includes clerical support staff and two individuals who recently were teachers in the same content area. Plan in some detail the meetings you will schedule on your first day at work.

This item involves **Level 3** complexity. The goal of planning meetings to be scheduled the first day is *stated broadly; inference is required to establish the problem* to be solved. For instance, the situations that will need to be addressed have not been given. Various people, if presented with this problem, would use *a number of different strategies* for planning the day. Any particular strategy would *involve multiple sequential and simultaneous steps or operations* to put together the first day's plan. Although previously learned strategies for planning these initial meetings would be employed, the specific *application of these strategies would likely be in a significantly different context* from that which occurred previously.

Initial Results from Using the Capabilities-Complexity Model

Initial data have been collected that provide useful information about the CCM. The data help identify test item qualities to which the model is sensitive. The data also indicate how well reviewers can use the model to classify test items. This section of the paper first describes the accuracy with which two groups of reviewers classified practice test items that were deliberately designed to measure cognitive qualities the model is intended to detect. Next, a series of analyses are presented related to reviewers' classification of items from two existing tests. The first set of items is from a subtest of the Florida Teacher Certification Examinations (FTCE) that measures the knowledge of prospective business teachers in the area of marketing. Six business educators who specialize in marketing and have been involved in FTCE item development reviewed these items. The second set of items is from the Grade 8 Science test of the Florida Comprehensive Assessment Test (FCAT). Three science educators who have similarly been involved in the development of middle-school science items for the FCAT reviewed those items.

Accuracy with Which Reviewers Classified Practice Test Items

The six marketing and three science educators reviewed a series of practice test items that were developed as part of the CCM training module. Immediately prior to viewing each set of items, the reviewers completed a training module through which they learned a particular skill relevant to using the CCM. The series of training modules covered these skills:

- distinction between items that measure declarative and procedural knowledge;
- distinction between problem solving, and declarative and procedural knowledge;
- levels of complexity for items that measure declarative knowledge;
- levels of complexity for items that measure procedural knowledge; and
- levels of complexity for items that measure problem solving.

Table 1 indicates how the marketing and science educators, following training, classified the practice test items related to the first skill: distinguishing between items that measure declarative and procedural knowledge.

Table 1
Classification of Items Measuring Declarative and Procedural Knowledge

	Marketing Educators		Science Educators	
	Declarative Knowledge	Procedural Knowledge	Declarative Knowledge	Procedural Knowledge
1. What is the meaning of the word <i>analogy</i> ?	6	0	3	0
2. Describe the characteristics of carbon dioxide.	6	0	3	0
3. Below are three sentences. Which one includes a simile? A. Snow is like a blanket, protecting plants from the cold. B. More rain is needed because of the extreme drought. C. The thunder scared everyone, coming so quickly after the lightning.	2	4	0	3
4. What is the Spanish word for <i>house</i> ?	6	0	3	0
5. Translate this French sentence into English: <i>Je vis dans une grande ville.</i>	0	6	0	3

Tables 2 through 5 similarly show how the educators classified the other sets of practice items. Shaded cells within each table identify the type of capability or level of complexity each test item was designed to measure. For instance, in Table 2, practice items 1 and 5 measure declarative knowledge; all six marketing and all three science educators correctly classified these items. Items 3 and 4 measure procedural knowledge and items 2 and 6 measure problem solving. Most or all of the marketing and science educators correctly classified these items.

Table 2
Classification of Items Measuring Declarative and Procedural Knowledge and Problem Solving

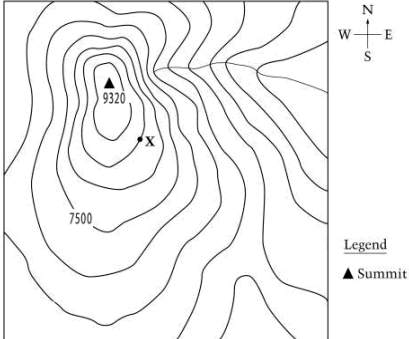

	Marketing Educators			Science Educators		
	Declarative Knowledge	Procedural Knowledge	Problem Solving	Declarative Knowledge	Procedural Knowledge	Problem Solving
1. <i>Describe</i> procedures used to amend the U.S. Constitution.	6	0	0	3	0	0
2. Develop plans for modifying a large residential home into a bed-and-breakfast lodging facility.	0	1	5	0	0	3
3. As a salesclerk, provide correct change to a customer who paid for a purchase.	0	6	0	0	3	0
4. When shown previously unseen jewelry that is representative of a particular archeological period, identify that period.	0	5	1	0	3	0
5. Explain why Earth's interior solid core, surrounded by an outer liquid core, contributes to its magnetic field.	6	0	0	3	0	0
6. Establish a plan for advertising a new, innovative cell phone.	0	2	4	0	0	3

Overall, the educators were able to independently correctly classify the practice test items. The distinction between procedural knowledge and problem solving occasionally caused some difficulty, as did the distinctions between some items that involved different levels of complexity. With the CCM, capabilities represent distinct categories; however, level of complexity is considered to be a continuum. For this reason, greater inconsistency was expected (and observed) in the classification of item complexity. These classifications represent each educator's first attempt at classifying test items using the model.

Table 3
Classification of Declarative Knowledge Items at Various Levels of Complexity

	Marketing Educators			Science Educators		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
1. What is the most common element in the universe? A. carbon B. hydrogen C. nitrogen D. oxygen	6	0	0	3	0	0
2. Explain how the human body protects itself when it gets too cold.	1	5	0	0	3	0
3. In physics, you learned that the screw and the inclined plane are two examples of types of simple machines. However, some regard them as two versions of the same machine. Why might they be considered the same machine?	0	2	4	0	2	1
4. With computers, RAM is the acronym for _____?	6	0	0	3	0	0
5. You learned that increased carbon dioxide in the atmosphere appears to be responsible for global warming, resulting in a significant decrease in the polar ice caps. If the carbon dioxide level immediately stopped increasing, why would the polar ice caps continue to melt?	0	2	4	0	0	3

Table 4
 Classification of Procedural Knowledge Items at Various Levels of Complexity

	Marketing Educators			Science Educators		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
<p>In this topographic map, each contour line represents 500 feet. The mountain summit is at 9320 feet. Point X is at what elevation?</p>  <p>Legend ▲ Summit</p>	2	3	0	2	1	0
<p>1. What is the most likely elevation at point X?</p> <p>A. 7500 feet B. 8000 feet C. 8500 feet D. 9000 feet</p>						
<p>2. This photograph of clouds was taken from an airplane. What type of clouds is pictured in the lower half of the photograph?</p>  <p>A. cirrus B. cumulus C. pileus D. stratus</p>	5	0	0	3	0	0
<p>3. Change the tense of the underlined verb to future perfect: I <u>finished</u> my lesson.</p>	4	1	0	3	0	0
<p>4. Compute the average of these five numbers: 2, 6, 4, 1, 7.</p>	0	5	0	0	3	0
<p>5. Purchase the items on your completed shopping list from a familiar grocery store.</p>	0	0	5	0	0	3

In all cases, the educators reached consensus in their classifications through discussion. With few exceptions, consensus was also reached when the educators later classified existing items from the FTCE and FCAT. There is always a danger that group dynamics will artificially bring about consensus. Certainly the potential for this occurring was present with the initial practice items since the person leading the training identified the expected response. This approach was necessary because these initial items represented a self-test and were an integral part of the training. With both groups of educators, however, dialogue was easy and coercion was not obvious. Consensus building when classifying actual items from tests appeared to involve comfortable dialogue, with no single individual dominating or shaping the discussion.

Table 5
Classification of Problem Solving Items at Various Levels of Complexity

	Marketing Educators			Science Educators		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
1. You are away on business for several days in a large city. This is your first trip to this city. You finished business early and want to attend a Major League Baseball game that will be playing that evening in the stadium several miles from your hotel. Tickets are available. How will you get there?	2	3	0	0	3	0
2. You are visiting a museum by yourself that you have not visited before. While you are on the ground floor in the museum, the fire alarm sounds and you smell smoke. No other people are nearby. How will you leave the building quickly?	1	4	0	3	0	0
3. You have not received your subscription to a weekly newsmagazine for two weeks. You want delivery to resume. What should you do?	2	2	1	1	1	1
4. Your only car has stopped working, and you discover it is beyond repair. What should you do?	0	1	4	0	0	3

Information in Tables 1 through 5 serves multiple purposes. As already noted, it suggests that differences in types of capabilities and levels of complexity can be detected in test items. The practice items also help clarify the qualities the CCM taps by providing additional examples of what would be considered declarative and procedural knowledge and problem solving, and also what represents each of the three levels of cognitive complexity. Data presented in these tables suggest that basic procedures of the CCM can be learned within the short period of time using the training modules developed for use with the model.

Application of the Capabilities-Complexity Model to Items from Existing Tests

As part of this study, the Capabilities-Complexity Model was used to classify existing items from the FTCE Marketing test and the FCAT Grade 8 Science test. With the exception of a subset of 20 FCAT items, these test items remain active within the respective tests and are discussed here in summary form only. The 20 released FCAT items are referenced more specifically in this paper.

As with the practice items, educators initially classified FTCE and FCAT items independently, and then reached consensus through discussion for any items with which there was a discrepancy in classification. Ideally, the model would result in full consistency, with each educator independently agreeing with the others with respect to the type of capability being measured and the level of complexity involved. In fact, were that level of consistency predictably obtained, there would be no need for using multiple educators when categorizing the test items. Nevertheless, consistency among raters is desirable. Tables 6 and 7 summarize the degree to which there was consistency.

Table 6 summarizes the consistency with which educators classified the type of capability being measured. The rows indicate the consensus classification that ultimately was assigned. Columns refer to the categorization independently assigned by the educators. The lower-level cell within the table indicates that, among FCAT items for which the ultimate consensus was that declarative knowledge is being measured, 152 of the independent ratings were in agreement with that eventual classification. In 24 instances, these declarative items were judged to measure procedural knowledge and in one instance to measure problem solving. Note that numbers within the cells refer to independent classifications across educators and test items.

Shaded cells within Table 6 refer to instances in which the independent judgments agreed with the eventual consensus judgment. The relative magnitude of numbers within the shaded cells suggests that, following training, educators are able to, with a high degree of consistency, independently establish the type of capability being measured by a test item.

Table 6
Consistency with Which Educators Independently Classified the Type of Capability Being Measured

		Individual Judgments with FCAT Items			Individual Judgments with FTCE Items		
		Declarative Knowledge	Procedural Knowledge	Problem Solving	Declarative Knowledge	Procedural Knowledge	Problem Solving
Consensus Ratings	Problem Solving	0	0	3	0	0	0
	Procedural Knowledge	11	51	4	18	58	2
	Declarative Knowledge	152	24	1	466	33	5

Table 7 similarly summarizes the consistency with which educators established the level of complexity involved. The rows indicate the consensus classification that ultimately was assigned. Columns refer to the complexity level independently assigned by the educators. The lower-level cell within the table indicates that, among FCAT items for which the ultimate consensus was that Level 1 complexity is present, 104 of the independent ratings were in agreement with that eventual classification. In 14 instances, the complexity was judged to be at Level 2 and in one instance at Level 3. As occurred when judging the type of capability, Table 7 suggests that, following training, educators are able to, with a high degree of consistency, independently establish the level of complexity of test items.

Table 7
Consistency with Which Educators Independently Classified the Level of Complexity Involved

		Individual Judgments with FCAT Items			Individual Judgments with FTCE Items		
		Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
Consensus Ratings	Level 3	0	0	0	0	0	0
	Level 2	10	60	2	16	41	2
	Level 1	104	14	1	203	24	0

Table 8 provides the number of FTCE and FCAT items judged to measure the three types of capabilities. In each set, the majority of test items—typically a large majority—were judged to measure declarative knowledge. This is not surprising, given that much of the knowledge one acquires is declarative in nature. However, the reason may also be that declarative items are easier to construct and include in large-scale assessments. Whether there is excessive emphasis on declarative knowledge is an important question—one that might better be addressed during the development of standards rather than speculated about during test item development and alignment. Only one test item was judged to measure problem solving. (The author believes this particular test item measures procedural knowledge, although the consensus of educators reviewing contradicts this opinion.) The lack of significant numbers of problem-solving items suggests an underrepresentation of this important type of capability within the assessments; however, it also reflects a limitation in the types of tasks that can be incorporated into large-scale assessments as presently conceptualized.

Table 8
Number and Percentage of FTCE and FCAT Items Judged to Measure the Three Types of Capabilities

	Number of Items			Percentage of Items		
	Declarative Knowledge	Procedural Knowledge	Problem Solving	Declarative Knowledge	Procedural Knowledge	Problem Solving
FTCE Live Items	84	13	0	86.6%	13.4%	0.0%
FCAT Live Items	47	14	1	75.8%	22.6%	1.6%
FCAT Released Items	12	8	0	60.0%	40.0%	0.0%

Oosterhof (in press) describes the importance of establishing the type of capability being assessed by constraining the cognitive process students must use in order to answer the test item. Merely looking at what the test item tells students to do is often insufficient for making this determination. Likewise, it is understandably important when developing test items (and standards) to use educators with extensive experience in working with students similar to those being assessed so that they will be sensitive to the mental process students will use. While using experienced educators is critical, it also is important that educators involved in item development and alignment studies be specifically trained in how to develop test items that constrain the cognitive process students will use to successfully answer the items.

Table 9 shows the number of items determined for each level of complexity for the declarative and procedural capabilities, and overall. The majority of test items were classified at Level 1 complexity, although a noticeably higher percentage of procedural rather than declarative items were at the higher level of complexity. This interaction between type of capability and level of complexity might be anticipated, although a legitimate question is whether items that measure declarative knowledge should heavily emphasize the lowest level of complexity, as appears to be the case with the present items. With declarative knowledge, differences between Levels 1 and 2 complexities include qualities like recalling facts and characteristics versus explaining properties and principles. In physics, for instance, should emphasis be placed on students being able to name and briefly describe the six simple machines (such as the inclined plane and the pulley), or should focus be on answering questions such as the following?

Why might a child who is unable to pull a wagon up a very steep hill be able to pull that same wagon up a less steep hill?

Asking students to briefly describe the six simple machines or asking them to explain differences in force required to pull the wagon uphill each measures declarative knowledge; however, the first involves Level 1 while the second involves Level 2. As noted, the present test items appear to heavily emphasize Level 1 complexity. One might anticipate a similar pattern on other large-scale assessments.

With respect to CCM criteria, no FTCE or FCAT test items included in the present study were classified as measuring Level 3 complexity.

Table 9
Number and Percentage of FTCE and FCAT Items Judged to Be at Each Level of Complexity

	Number of Items			Percentage of Items		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
FTCE Live Items						
Declarative Items	35	5	0	87.5%	12.5%	0.0%
Procedural Items	7	6	0	53.8%	46.2%	0.0%
Overall	42	11	0	79.2%	20.8%	0.0%
FCAT Live Items						
Declarative Items	22	8	0	73.3%	26.7%	0.0%
Procedural Items	8	6	0	57.1%	42.9%	0.0%
Overall	30	14	0	68.2%	31.8%	0.0%
FCAT Released Items						
Declarative Items	8	4	0	66.7%	33.3%	0.0%
Procedural Items	2	6	0	25.0%	75.0%	0.0%
Overall	10	10	0	50.0%	50.0%	0.0%

(Note that Tables 8 and 9 involve different numbers of test items. For instance, Table 8 shows that 84 FTCE items were judged to measure declarative knowledge, while Table 9 involves 40 FTCE items [35 at Level 1 and 5 at Level 2] with declarative knowledge. This discrepancy relates to how data for the study were gathered. With both the FTCE and FCAT tests, items were oversampled in preparation for the study to ensure that the sample would not be exhausted during the study. The six marketing educators and three science educators, who traveled from distant parts of Florida for their respective sessions, were available to the present study for one day. Approximately half this period was scheduled for identifying the type of capability measured by the sampled items. When the scheduled time was exhausted, the next step was to establish the level of complexity for the items that had been classified. Because of their smaller number, the subset of procedural items was oversampled during this next step to ensure an adequate number of procedural items with which to include in the analysis of complexity. As anticipated, when the scheduled amount of time was exhausted, the complexity level had not been determined for all available items. Tables 8 and 9 report the numbers of test items that were classified at each stage within the process.)

Table 10 lists the item numbers for the FCAT released items that were included in this study, and also provides the consensus capability and complexity that were established for each item. To conserve space, these Grade 8 Science test items are not included in this paper, but they are available for review on the FDOE Web site.⁴

Table 10
FCAT Grade 8 Science Released Items Included in Study

Item Number	Type of Capability	Level of Complexity
1	Declarative	Level 2
2	Procedural	Level 1
3	Declarative	Level 1
7	Procedural	Level 2
8	Declarative	Level 1
14	Declarative	Level 1
15	Procedural	Level 1
16	Procedural	Level 2
21	Procedural	Level 2
22	Declarative	Level 1
24	Declarative	Level 1
29	Procedural	Level 2
30	Declarative	Level 2
33	Procedural	Level 2
35	Procedural	Level 2
36	Declarative	Level 1
38	Declarative	Level 2
39	Declarative	Level 2
41	Declarative	Level 1
42	Declarative	Level 1

The complexity of many but not all of the FCAT live and released items had been previously established by FDOE using an adaptation of Webb’s DOK model. This adaptation is described in an FDOE Web site.⁵ As with the CCM, FDOE’s adaptation of the DOK model involves three levels of item complexity. Currently, FDOE refers to these complexity ratings as *content difficulty*, although it recognizes that item complexity—and not necessarily item difficulty—is involved. Table 11 is a cross-tabulation of the 61 live FCAT Science items for which both FDOE and CCM complexity ratings are available. For instance, 8 of the 61 items were classified at Level 2 by both rating systems. If the two systems were congruent, all items for which both complexity ratings are available would fall within the cells of the shaded diagonal of the table.

Table 11
Comparison of Complexity Ratings

		FDOE Content Difficulty Ratings (Complexity)				
		1	2	3	Unknown	Totals
CCM Consensus Complexity	3					0
	2		8	2	4	14
	1	8	18		4	30
	Not Rated	10	4	1	2	17
	Totals	18	30	3	10	61

The complexity component of the CCM was derived through a review of adaptations of Webb’s DOK model that have been made publicly available by various state-level departments of education. Unlike the other adaptations, the CCM defines complexity in light of the type of capability involved. The CCM also uses generic descriptions for each complexity level, whereas Webb has developed specialized descriptions for different content areas such as science, mathematics, and reading.

Tables 12 and 13 list p-values (proportion of students answering an item correctly) for FCAT live items for which a complexity rating was available. Table 12 associates p-values with content difficulty (FDOE’s adaptation of Webb’s model); Table 13 does the same for the CCM. In both tables, p-values are ordered from high to low to facilitate interpretation. Tables 12 and 13 are not directly comparable because each is based on somewhat different test items. In both cases, however, item complexity clearly is concerned with something other than the difficulty of test items. As noted earlier, some highly complex tasks (such as purchasing a book from an online retailer) are easy to complete despite their complexity. Other tasks, such as recalling obscure facts, are cognitively much less complex even though they can be quite difficult to answer correctly. Thus, increasing the complexity of tasks included in an assessment is not equivalent to making the test more difficult.

Table 12
Observed p-Values for FCAT Middle School Science Items of Various Content Difficulties

Content Difficulty	Observed p-Values
3	.59 .16 .16
2	.85 .84 .82 .80 .76 .76 .75 .74 .73 .73 .71 .69 .68 .68 .67 .66 .63 .61 .61 .59 .56 .56 .51 .50 .50 .50 .49 .46 .44 .39 .33
1	.84 .82 .74 .70 .69 .69 .66 .65 .61 .59 .59 .58 .54 .54 .52 .49 .48 .46

Table 13
Observed p-Values for FCAT Middle School Science Items of Various CCM Complexity Levels

CCM Complexity	Observed p-Values
3	
2	.75 .72 .69 .68 .63 .61 .59 .59 .58 .49 .46 .39 .16
1	.84 .82 .80 .78 .76 .76 .74 .73 .73 .70 .69 .68 .67 .66 .66 .66 .63 .61 .61 .59 .56 .56 .54 .51 .48 .48 .46 .44 .33 .22

Summary and Conclusions

The CCM classifies test items by type of cognitive capability (declarative knowledge, procedural knowledge, and problem solving) and level of cognitive complexity. Criteria used to determine the level of complexity vary, depending on the type of capability involved.

Types of capabilities and levels of complexity have represented important considerations in assessment for some time. For example, although Bloom’s taxonomy has been widely used to identify types of learning outcomes, subcomponents included in each category of the taxonomy exhibit a wide range in levels of complexity. However, accommodating levels of complexity within a classification system does not ensure that levels of complexity are appropriately incorporated into important assessments. Webb’s DOK model addresses this issue and explicitly identifies qualities of different complexity levels. The CCM builds on Webb’s model and proposes complexity criteria for different types of capabilities. The CCM replaces Bloom’s taxonomy with the more contemporary categories of declarative knowledge, procedural knowledge, and problem solving.

To use the CCM, one first establishes the type of capability involved and then, based on that classification, applies criteria for classifying the level of complexity. This two-step process allows the classification of complexity to be more focused. Anecdotal evidence suggests this two-step sequence simplifies the process, perhaps because item reviewers focus on specific attributes of complexity relevant to a particular type of capability. This paper describes the CCM and presents results associated with its initial application.

Training modules were developed to facilitate use of the CCM. These modules were used with two groups of educators when they employed the CCM to classify FTCE and FCAT test items. With the training, the educators were able to reach consensus on the types of capabilities test items measured as well as their level of complexity.

An interesting question not investigated in the present study is whether the CCM, with its two-step process that separates type of capability and level of complexity, results in higher rater agreement brought about by the more focused attention to complexity attributes associated with respective types of capability. Webb, Herman, and Webb (2006) suggest that inconsistency between raters may be a significant concern within test alignment studies. If inconsistency among raters is a problem, this would noticeably attenuate correlations between educational standards and test items used to assess student achievement of those standards. Increasing rater agreement would represent an important approach to addressing this attenuation.

Both type of capability and level of complexity are highly relevant to the development of assessments. Explicitly addressing both seems prudent. The science educators involved with the present study stated emphatically that examining both the types of capabilities and levels of

complexity helped them understand why students had difficulty learning particular content and performed poorly on selected test items when it appeared that test preparation had been adequate.

Inasmuch as types of capabilities and levels of complexity should be explicitly controlled during the development of important assessments, it seems reasonable that they also should be unambiguously specified as part of educational standards and benchmarks. Doing so could improve the test alignment process. It also would help delineate parts of the curriculum that involve problem solving and cognitively complex aspects of declarative and procedural knowledge, areas the present investigation illustrates are underrepresented in large-scale assessments. Establishing what assessments can and cannot do well represents an important consideration for consequential evidence of validity. Clearly conveying types of capabilities and levels of complexity in educational standards, particularly in benchmarks or other delineations of standards, could help establish the proper limits of large-scale assessments. This should also facilitate a cautious interpretation and use of these assessments. It may also encourage the meaningful inclusions of other assessments, including those of teachers, to evaluate proficiencies involving important types of skills that are out-of-reach of conventional large-scale assessments.

References

- Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives, handbook 1: Cognitive domain*. New York: McKay.
- Confrey, J. (1990). A review of the research on student conceptions in mathematics, science, and programming. In C. B. Cazden (Ed.), *Review of research in education: Vol. 16* (pp. 3–56). Washington, DC: American Educational Research Association.
- Gagné, E. D, Yekovich, C. W., & Yekovich, F. R. (1993). *The cognitive psychology of school learning*. New York: Harper Collins.
- Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematics test with Bloom's taxonomy. *Journal of Educational Research, 91*, 26–32.
- Oosterhof, A. (in press). *Developing and using classroom assessments* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Snow, R. E. (1989). Towards assessment of cognitive and conative structure in learning. *Educational Researcher, 18*(9), 8–14.
- Tittle, C. K., Hecht, D., & Moore, P. (1993). Assessment theory and research for classrooms: From taxonomies to constructing meaning in context. *Educational Measurement: Issues and Practice, 12*(4), 13–19.
- Webb, N. L. (2002). *Depth-of-knowledge levels for four content areas*. Retrieved from <http://facstaff.wcer.wisc.edu/normw/All%20content%20areas%20%20DOK%20levels%2032802.doc>.
- Webb, N., Herman, J., & Webb, N. (2006). *Alignment of mathematics state-level standards and assessments: The role of reviewer agreement*. CSE Technical Report 685. Los Angeles: Center for the Study of Evaluation, University of California. Retrieved from <http://www.cse.ucla.edu/products/reports/R685.pdf>.

Footnotes

¹See http://www.paec-fame.org/reading_docs/RDGMATHHPD07PAEC.pdf, p. 4.

²See <http://professionaldevelopment.brevard.k12.fl.us/documents/induction/wise%20training.pdf>, p. 83.

³The text on pages 5–16 of this report appeared previously in *The Capabilities-Complexity Model Handbook*, which was developed by the Center for Advancement of Learning and Assessment for the Florida Department of Education. This material is copyrighted by the State of Florida, Department of State © 2007. All rights reserved. It may not be reproduced in any form without the express written consent of the Florida Department of Education.

⁴The FCAT Grade 8 Science test items released fall 2007 are available at http://fcats.fldoe.org/pdf/releasepdf/07/FL07_G8S_TB_Rel_WT_C001.pdf.

⁵See http://fcats.fldoe.org/pdf/cog_complexity-fv31.pdf.