# Designing a Cognitive Skills Framework for Item Development

Rebecca J. Kopriva
University of Wisconsin

Phoebe C. Winter
Pacific Metrics Corporation

December 3, 2012

This work was supported by a Grant for Enhanced Assessment Instruments awarded by the U.S. Department of Education to the Virginia Department of Education as lead state, award S368A090029. The opinions in the document are those of the authors and do not necessarily reflect those of the North Carolina Department of Public Instruction, the U.S. Department of Education, or the participating project states.

# Contents

# Section 1: Introduction

Practitioners in content areas in K-12 education have focused on describing knowledge, reasoning, and cognitive skills they believe students should be able to demonstrate in order to be considered proficient. In recent years, science education has been a particularly fertile area for this work, as curriculum and instruction have moved towards defining scientific thinking and emphasizing the need for students to understand the scientific process as well facts and theories (e.g., Duschl, Schweingruber, and Shouse, 2007; College Board, 2011; National Research Council, 2011). A number of researchers have worked with the cognitive literature and content frameworks such as these to provide a principled basis from which to define the academic domains, for purposes of specifying what should be measured in assessments, and then for evaluating the fit of items with the proposed construct specifications or intentions (e.g., Baxter and Glaser, 1997; Shavelson, Ruiz-Primo, Li, and Avela, 2003; Graf, Peterson, Lawless, Steffen, 2005; Li, Ruiz-Primo, and Shavelson, 2006; Ruiz Primo, 2009 ). In several cases, suggestions have been made about the types of assessment items or tasks that might be used to provide evidence of these valued claims. However, there appears to be a dearth of literature about exactly how to operationalize the items or tasks in order to produce the specific evidence about student proficiency outlined by the various explanations of targeted knowledge, reasoning or skills.

The purpose of this project is to extend the work begun by researchers such as Shavelson, Ruiz-Primo and others, and to consider if and how we might move towards a set of parameters of definable, manifest characteristics of items or tasks that can reliably produce outcomes we can interpret as being consistent with our valued inferences. That is, we are attempting to connect the latent domains associated with specific inferences with manifest features of tasks developed to support those inferences. Our work draws on and is intended to extend current related work in more precisely specifying features of items and tasks. For instance, Luecht (2009) has identified three item feature

continua, each of which vary by complexity. By keeping content specifications the same but changing the position of each feature along the continua, he argues that new items can be built to measure more or less cognitively complex construct targets or new items that measure the same construct but vary in difficulty. His item feature continua are examples of the types of characteristics that may be key in distinguishing valued performance. We are evaluating work such as this, in order to begin to determine how the most salient manifest features might be best combined and best 'manipulated' to produce desired latent outcomes.

This project focuses on only one aspect of item development – constructing items to target specified cognitive skills -- although it is a foundational one. There are other very important considerations that are part of the item development process that are not addressed in our work. For example, assuring that items and task access the knowledge and skills of the entire target population, including student with disabilities, English learners, and other students who may have differing access needs is not a focus of our work. We expect that our work will be used in conjunction with other critical considerations that enter into the item development process.

## Relationship to Evidence Centered Design

Evidence centered design (ECD) provides a framework for item and test development that explicitly connects item and task development to the claims the test user wants to make about examinee knowledge, skills, and abilities. To do this, ECD suggests a principled method of moving through five "layers" of test development (Mislevy & Haertel, 2006): (1) domain analysis, which involves collecting information about the targeted domain, including how it is taught and learned and how concepts are communicated; (2) domain modeling, specifying what the assessment is supposed to measure on the form of an assessment argument; (3) conceptual assessment framework, which is essentially the task and assessment development model, encompassing task features, specifications, and measurement

models; (4) assessment implementation, including the development of tasks, finalization of scoring rules, and test assembly; and (5) assessment delivery, the administration, scoring and reporting functions. The ECD process has been used in research (e.g., Alves, Gierl, & Lai, 2010) and its principles are beginning to be implemented in operational test development, most notably in the development of a number of Advanced Placement Program tests (Huff, Steinberg, & Matts, 2010) and as a foundation for the development of the Smarter Balanced Assessment Consortium tests (Measured Progress/ETS Collaborative, April 16, 2012). We believe that our work can strengthen applications of ECD principles by guiding item writers in the specifics of how to support various levels of claims.

## Section 2: Overview of the Literature

The 1993 edited volume, *Test Theory for a New Generation of Tests* (Frederikson, Mislevy, & Bejar) was based on the premise that cognitive theory should inform how we develop and interpret the results of tests. In 2001, Knowing What Students Know: The Science and Design of Educational Assessment (National Research Council) reiterated that message and provided a foundation for research and test development that is manifest in recent test design, from Mislevy and colleagues Principled Assessment Designs for Inquiry (cf., Mislevy, Steinberg, Almond, Haertel, & Penuel, 2003; Seeratan& Mislevy, 2009) to the Smarter Balanced Assessment Consortium's use of evidence centered design principles in structuring their assessment framework (Measured Progress/ETS Collaborative, April 16, 2012). Parallel work on structuring curriculum according to how students learn is exemplified by the National Research Council publication on science education, which moves curriculum and instruction towards defining scientific thinking and emphasizes the need for students to understand the scientific process as well facts and theories (National Research Council, 2011). In our work, we reviewed selected literature on problem solving, cognition, and task design to explore the specific aspects of academic tasks that were

thought to require the application of specific types of knowledge, skills and abilities, affecting how the tasks function. A sampling of the work we reviewed is discussed here.

In recent years, science education has been a particularly fertile area for this work, as curriculum and instruction have moved towards defining scientific thinking and emphasizing the need for students to understand the scientific process as well facts and theories (e.g., Duschl, Schweingruber, and Shouse, 2007; College Board, February 2011; National Research Council (NRC), 2011). NRC's (2011) science education framework emphasizes the importance of problem-solving and scientific inquiry processes to each of the core disciplines in the framework, connecting argumentation and analysis skills to the development of knowledge in science.

Jonassen and Hung (2008) drew from the problem-solving literature to define a number of factors that affect problem difficulty in the context of problem-based learning along two domains, complexity and structure. A problem's complexity is influenced by the breadth of knowledge required to solve the problem, the degree of abstraction of the concepts involved, the number of steps and the intricacy of the steps needed to solve the problem, and the number of relationships that need to be processed to address the problem. Problem structure is influenced by the extent of the unknowns in the problem (intransparency), the number of possible interpretations of the problem, the degree to which the problem requires interdisciplinary knowledge, the degree to which dynamic variables are involved, and the number of legitimate solution pathways. The factors that define problem complexity and structure provide some initial guidance into how the structure of assessment tasks may affect the types of knowledge, skills, and abilities that are needed to appropriately respond to the tasks.

In research closer to the assessment realm, we reviewed literature on learning progressions. In developing learning progressions, researchers examine the increases in the types and content of knowledge that students demonstrate as they move towards a deeper understanding of the concept

(National Research Council, 2006). Developing learning progressions requires an explicit use of research and expert judgment on how students learn, and the collection of empirical data that is employed to adjust the learning progression. The resulting models are used to evaluate how students are progressing in their learning of the identified concept and to determine what the next step is in helping the student reach the desired level of learning. For example, Kennedy and Wilson (2007) created a learning progression for middle school science on the topic of buoyancy. They mapped out specific knowledge that the student had to attain and levels of reasoning that changed as the student progressed in their understanding of buoyancy along eight levels. Figure 1 shows the reasoning portion of progression that Kennedy and Wilson developed (2007, page 37).

Figure 1. Reasoning portion of learning progression on buoyancy – extracted from Kennedy and Wilson (2007).

| Level | What the Student Can Already Do | What the Student Needs to Improve |
|---|---|---|
| P | Principled<br>Student uses an explicit principle that applies to objects in general.<br>"An object floats when its mass is large." | |
| R | Relational<br>Student uses a specific relationship in which the object, the property, and the magnitude of the property (e.g., more vs. less) are all clear.<br>**Note:** Some of the parts of the relationship may be made clear by the item stem, or by another part of the response (e.g., a prediction), rather | To progress to the next level, student needs to use a principle that would apply to objects in general. |
| U | Unclear Relational<br>Student uses a specific relationship in which either the object, the property, or the magnitude of the property (e.g., more vs. less) is not clear. | To progress to the next level, student needs to explicitly identify all three parts of the relationship in their explanation. |
| E | Experiential<br>Student justifies their answer by appealing to prior experience, in the form of a personal observation or an authoritative source.<br>"It floats because that's what we saw in | To progress to the next level, student needs to use a relationship to explain their answer, not just evidence to justify it. |
| IE | Inadequate Explanation<br>Student either restates their answer as an explanation, or simply asserts that their answer is correct. | To progress to the next level, student needs to understand what evidence is and the relationship between evidence and an |
| OT | Off Target<br>Student cannot or does not give an explanation for their answer.<br>"I have no idea." | To progress to the next level, student needs to justify their answer in some way. |
| NR | No Response<br>Student left the response blank. | To progress to the next level, student needs to respond to the question. |

Gotwals and Singer (2009) developed a framework that describes what they termed "middle knowledge". This middle knowledge represents the intermediate steps to reaching a sophisticated understanding of a topic, rather than focusing on end-stage knowledge tapped by most educational assessment. Their framework, shown in Figure 2, links student response features to levels of reasoning and content knowledge,

Figure 2. Content-reasoning matrix – extracted from Gotwals and Singer (2009, page 265).

**Content-Reasoning Matrix for Tasks Associated With Constructing Evidence-Based Explanations About a Key Science Topic**

| | | Scientific Content | | |
|---|---|---|---|---|
| | | Minimal Content | Intermediate Content | Complex Content |
| Scientific Reasoning | Minimal Reasoning | Students are given evidence and claim. **Students match relevant evidence to a given claim.** Minimal content knowledge is required. | Students are given evidence and claim. **Students match relevant evidence to a given claim.** To do this, students must apply a moderate amount of additional (not given) content knowledge. | Students are given evidence and claim. **Students match relevant evidence to a given claim.** To do this, students must apply a substantial amount of additional (not given) content knowledge. |
| | Intermediate Reasoning | **Students construct a simple explanation using prompts or partially completed explanations to guide, for example, the development of a claim and the use of relevant evidence.** Minimal content knowledge is required. | **Students construct a simple explanation using prompts or partially completed explanations to guide, for example, the development of a claim and the use of relevant evidence.** Construction draws on a moderate amount of content knowledge. | **Students construct a simple explanation using prompts or partially completed explanations to guide, for example, the development of a claim and the use of relevant evidence.** Construction draws on a substantial amount of additional (not given) content knowledge. |
| | Complex Reasoning | **Students construct a scientific explanation without any prompts or guidance.** Minimal content knowledge is required. | **Students construct a scientific explanation without any prompts or guidance.** Construction draws on a moderate amount of content knowledge. | **Students construct a scientific explanation without any prompts or guidance.** Construction draws on a substantial amount of additional (not given) content knowledge in order to, for example, determine salient from irrelevant evidence and to justify claim through scientific reasoning. |

A number of researchers have concentrated on identifying aspects of items and tasks that affect how the items and tasks function. This body of research provided the source for the iterations of our model as described in the next section.

## Item and Task Features

Feltovich, Spiro, and Coulson (1993) describe dimensions that affect the complexity of tasks and concepts, arguing that misconceptions are more likely to be adopted and are more stable when they fall on the simpler ends of each of the continua. These dimensions are relevant to designing tasks that target particular levels of sophistication in knowledge, skills, and abilities: concreteness/abstractness, or the degree to which the concepts and processes involved are tangible; discreteness/continuity describes whether the processes in the task are distinct from each other; static/dynamic is the degree to which the necessary processes change depending on context; sequentiality/simultaneity refers to whether the processes used are applied one at a time, in sequence, or whether multiple process must be employed

at the same time; mechanism/organicism reflects the degree to which the process are mechanistic, attributable to an identifiable agent, or holistic in nature; universality/conditionality is the degree to which the principles applied to the problem are general in nature or whether they are conditional on the context of the problem; and linearity/nonlinearity describes the relationships among the processes employed. The continua are not independent of each other, but they do differentiate between the types of characteristics a task might have that may affect the types of knowledge and skills required to successfully address the task.

More recently, a group of researchers have investigated methods for explicitly connecting cognitively-based content domain descriptions to what should be measured in academic assessments, mostly in the science domain. In research that has set the stage for additional work, Baxter and Glaser (1997) characterize science tasks by their joint location on two continua that describe the task requirements: lean to rich science content knowledge and constrained to open science process skills. These categories have been further refined by subsequent researchers, most prominently Ruiz-Primo and her colleagues (Shavelson, Ruiz-Primo, Li, and Ayala, 2003; Li, Ruiz-Primo, and Shavelson, 2006; Li and Tsai, 2007; Ruiz-Primo, 2009).

In Shavelson et al. (2003) and Li, Ruiz-Primo, and Shavelson (2006), the authors describe four categories they used to describe item and task requirements to characterize the content of the 1999 Trends in International Math and Science Study, building on Baxter and Glaser's earlier work: task demands, inferred cognitive demands, item openness and complexity. The definitions are shown in Figure 3 (Shavelson et al., 2003, page 10).

Figure 3: Item Coding Categories for Science Assessment, extracted from Shavelson, et al. (2003)

**Task Demands:** What does the item ask examinees to do?

- Terms, symbols, vocabulary, and definitions
- Statements, descriptions, and facts
- Procedures, steps, and actions
- Algorithms and equations, figures and diagrams, and tables
- Models, relationships, theories, explanations, and principles

**Cognitive Demands:** Inferred cognitive processes examinee may bring to task and how they use knowledge and reasoning to respond to the item.

- Visualization
- Mathematical calculation
- Mechanical operation (e.g., draw diagram, balance chemical equation)
- Perform experiment
- Recall information/fact
- Reason and interpret with models and principles
- Describe and record information and outcomes
- Select and use strategies
- Plan and monitor
- Guess or eliminate wrong options
- Reason with common sense

**Item Openness:** Degree of freedom student has in shaping response to item.

- Hands-on vs. paper-and-pencil
- Read options and choose vs. generate responses on own (selected vs. constructed-response format)
- Require only information found in task vs. steps/information can be learned from the task
- Require one vs. multiple correct solutions/approaches
- Follow instructions or steps

**Complexity:** Familiarity, relevance, reading difficulty, common experience of the item

- Textbook-type task vs. ill-structured task (provides or contains new situation/information)
- Inclusion of irrelevant background information
- Long, reading demanding descriptions and complicated vocabulary
- Answers contradict every experience/belief

Li and Tsai (2007) expanded on this work, going further in tying the categories to types of knowledge required by the items: declarative, procedural, strategic, and schematic. They evaluated TIMSS and several state tests according to the framework and found that there was a relationship between type of knowledge and their coding categories.

In 2009, Ruiz-Primo produced a comprehensive monograph for the National Resource Council that explored different but related models of items and the inferences one can make based on item types, with the goal of developing and evaluating assessments measuring 21st century skills (Ruiz-Primo, 2009). The author proposes that an analysis of items focus on three dimensions: task demands, cognitive demands, item openness, and item complexity. Figure 4 illustrates these dimensions (Ruiz-Primo, 2009, page 64).

Figure 4. Item and Task Dimensions for Logical Analysis, extracted from Ruiz-Primo (2009)

| Task Dimensions | Examples for Designing Tasks or for Coding Decisions | | | |
|---|---|---|---|---|
| **Task Demands:** What the task asks the test taker to do, what it elicits from the student | - Defining concepts<br>- Identifying facts | - Executing procedures in familiar tasks<br>- Executing procedures in unfamiliar tasks | - Selecting an appropriate procedure to solve a problem<br>- Determining the theoretical position of a manuscript<br>- Drawing diagrams illustrating a process<br>- Writing an equation from a statement<br>- Constructing an interpretation<br>- Drawing conclusions<br>- Justifying or predicting | - Evaluating the validity of a conclusion<br>- Evaluating products, or proposals<br>- Producing alternative solutions to a given problem<br>- Designing an experiment to solve a non-routine or novel problem |
| **Cognitive Demands:** Inferred cognitive processes that students likely act upon to provide responses | Less Cognitive Demanding ← → More Cognitive Demanding | | | |
| | *Remembering*<br>Recognizing<br>Recalling | *Applying*<br>Executing or Implementing more or less routine procedures | *Reasoning Using Mental Models*<br>Explaining<br>Interpreting<br>Inferring<br>Organizing/Classifying<br>Comparing/Contrasting<br>Exemplifying | *Assembling Knowledge in New/Creative Ways*<br>Planning<br>Generating<br>Producing<br>Monitoring |
| **Item Openness** Gradation in the constraint exerted in the nature and extent of the response | - The continuum from multiple-choice tests to constructed response test format such as long essays, projects, and collection of products over time<br>- Require one correct solution versus multiple correct solutions/approaches<br>- The continuum of the structureness/directedness of the task (following instructions or steps) | | | |

In a related line of research, Ferrara and his colleagues (Ferrara et al., 2004; Ferrara and Duncan, 2011) developed and implemented a framework used to characterize 6th grade science assessment items. The authors used Item Demands Analysis (IDA) coding to compare to the achievement targets in the test specifications document to items on the test. The definitions used by Ferrara, Huff, and Lopez (2010) to characterize cognitive complexity are shown in Figure 5 (page 6).

Figure 5. Characterizing Task Complexity, extracted from Ferrara, Huff, and Lopez (2010).

| | Cognitive complexity increases when the target of measurement: |
|---|---|
| Dissonance | • carries misconceptions and preconceptions that are untested |
| Multidimensionality | • involves changes in three-dimensional structure<br>• requires multi-step methodologies or mathematical processes<br>• *involves the molecular level*<br>• involves simultaneous equations<br>• *involves simultaneous conceptualization of multiple variables*<br>• involves multiple, possibly confounding causes of an effect |
| Abstractness | • involves models that approximately represent a phenomenon<br>• requires integration of factors at very different scales<br>• *requires weighing and assessing of relative causal significance*<br>• *involves multiple data sources*<br>• *involves heterogeneous data*<br>• *involves selection and representation of variables*<br>• *involves mathematical representations*<br>• involves conceptual statistics including precision and significance |
| Contingency | • *involves multiple competing criteria*<br>• *requires flexible generalization of a heuristic*<br>• *involves multiple variables whose significance must be decided*<br>• *involves randomness*<br>• *involves fuzziness of criteria in categorization*<br>• involves the possible confusion of correlation and causation |
| Transfer | • novel concepts that combine existing concepts<br>• novel situations that require extension of existing knowledge |

Luecht (2009) has also attended to task and item features, but to manipulate item difficulty rather than to explain cognitive factors needed to respond to the item. Luecht defines task models and related templates along four dimensions that can be varied to affect task difficulty without affecting the specific

construct being assessed. Luecht writes that three of the dimensions are cognitively based: (1) task complexity, the number and complexity of targeted construct-related skills and actions needed to complete the task; (2) information density, the concentration of text and graphics; and context complexity, partially defined by the presence of distractive information. The fourth dimension Luecht defined is facility and utility of auxiliary tools and information, referring to the availability of tools such as calculators for examinees to use when addressing a task. Luecht's approach of varying items along specific dimensions to affect task difficulty, is similar in principle to the approach we have taken to producing items that are likely to address identified assessment targets.

The research with a purpose most closely aligned to our goals was conducted by Oosterhof, Rohani, Sanfilippo, Stillwell, and Hawkins (2008). Oosterhof et al. developed a model defining features of tasks and items specific to each of three type of knowledge (declarative, procedural, and problem-solving) that affected the complexity of items measuring that type of knowledge. For example, Oosterhof et al. posited that the complexity of a task measuring procedural knowledge is affected by five factors: (1) number of steps involved in the procedure used to address the task; (2) the degree to which directions are direct or require an inference on the part of the examinee; (3) the degree of abstraction involved; (4) whether the procedures constrain the variability of the actions required to address the task; and (5) the similarity between the item context and the context in which the required procedure was learned. In a small study involving marketing and science items, the authors found that judges could consistently classify items according to knowledge type and rate complexity consistently, using definitions that incorporated the defined item features. As shown in Figure 6, the categories of item features identified were different for each type of knowledge – that is, the categories were not consistent across type of knowledge: for example, number of steps or operations was identified as a variable affecting procedural knowledge tasks but not declarative knowledge tasks. This limits the direct utility of the model for developing items and tasks.

Figure 6. Item features by type of knowledge -- extracted from Oosterfhof, et al., 2008.

Declarative Knowledge

| 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. |

Procedural Knowledge

| 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. |

Problem Solving

| 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. |

## Section 3: Developing a Framework for Guiding Item Writers

Clearly, a number of learning theory writers have explored in detail the connections between aspects of how knowledge, skills and abilities are learned and prominent cognitive stimuli that are activated as part of learning. The literature we reviewed also illustrated a keen interest in considering how to link what we want to measure and the inferences we want to draw with how to distinguish test products that produce these intended outcomes. Some literature focused on identifying relevant evidence attached to specific valued content and cognitive skills; some focused on identifying critical aspects of items that appear to activate particular cognitive schemas at various level of complexity; and some focused on item components that seem to more often be connected to certain types of evidence. Some authors developed more or less sophisticated item evaluation categories that parsed the elements of items and described post hoc how these elements may contribute to producing intended inferences. Old item generation methods focused on decreasing or increasing item difficulty by replacing more or less basic to challenging content knowledge into item shells, and literature and practice have postulated how different item types might contribute to expanding the range of cognitive complexity we might be able to measure.

This repository provides a fertile foundation of descriptive literature upon which to consider how to more precisely guide writers in producing items that yield evidence directly linked to intended outcomes that extend beyond just changes in difficulty. The work done to date has yet to produce an explicit, systematic approach to item generation that can be reliably exploited by item writers. In part this is because the procedure has to be both user-friendly and yield items and tasks that faithfully replicate desired results. Further, to be most useful, the approach should be applicable across different standards within a content area and generalizable across content areas. Too much detail and complexity in the method encourages confusion; too little complexity invites error-prone results. As the last section summarizes, Shavelson and his colleagues (Shavelson, Ruiz-Primo, Li and Ayala, 2003; Ruiz-Promo, 2009;

Li and Tsai, 2007) have explored the criteria underlying the specification of item features. Luecht (2009) has tackled the charge of broadening automated item generation approaches by identifying three item feature continua, each of which vary by complexity. We see that each of these lines of work have tremendous potential.

## Overview of the Process

The project began with a kick-off meeting in February 2011 with a panel of experts. A list of the panel members can be found in Appendix I.2. A literature review had been completed for this meeting, and this review was discussed at the meeting (Appendix I.8). Additionally, several items were presented and a discussion followed to brainstorm about the aspects that differentiated items and why. All in all, it became clear that an extended examination and discussion of the literature was needed, and that more focused work needed to be done by a small group of seasoned assessment members.

A subgroup of the panel was identified. These members were the PI and co-PI, Rebecca Kopriva and Phoebe Winter, respectively, Doris Redfield, President of Edvantia, and Sue Rigney, ex-officio member from USED. The charge to the subgroup was to think about how to approach the task of providing specific, concrete literature-based direction to item writers writing different kinds of items. The direction would simultaneously recognize the diversity of evidence across and within typical content/complexity cells, recognize that different types of questions yield different types of valued evidence, and recognize that item features target and constrain the types of evidence that is collected. For the guidance to be differentially useful to writers tasked with measuring different kinds of knowledge and skills, the direction would seek to link the diverse intents of items to particular criteria such that the details of the criteria would ideally differ across some type of profiles. Each of these would reflect unique direction not replicated elsewhere, but would not be overly prescriptive or cumbersome to implement.

This subgroup combed the literature and met five times from June, 2011 through January of 2012, with reviews and work in between. A webinar with panel members was convened in August 2011 to discuss the work so far, and individual panel members were contacted as relevant to provide feedback. Many attempts at conceptualizing the features of a possible framework were considered by the subgroup and then refined or put aside. The approach the authors have taken throughout this project has been to iteratively propose possible candidates, evaluate each using available items, and build on or re-conceptualize the method as relevant. The framework presented here is the culmination of the work to-date.

## I. Item Development Framework: Conceptualization

The framework, conceptualized as a two-tiered model, is depicted in Figure 1. Set within the test domain articulated by the content standards and the test specifications that identify particular standards, objectives or topics the test will cover, are the assessment targets for each standard or objective. These targets unpack the complexity dimension of the test specifications, asking developers to select which cells in the matrix they are intending items to be written within.

The second tier of the framework model is designed to identify item writing characteristics that items must have in order to properly measure each assessment target cell. A number of item writing features and their levels have been identified as central to the item writing process. As a set, the relevant levels of the features in each cell will stipulate what items must contain to produce the evidence intended by the test developers. The framework to date is conceptualized to be broadly applicable across subjects and across standards or some type of expression of valued content within subjects.

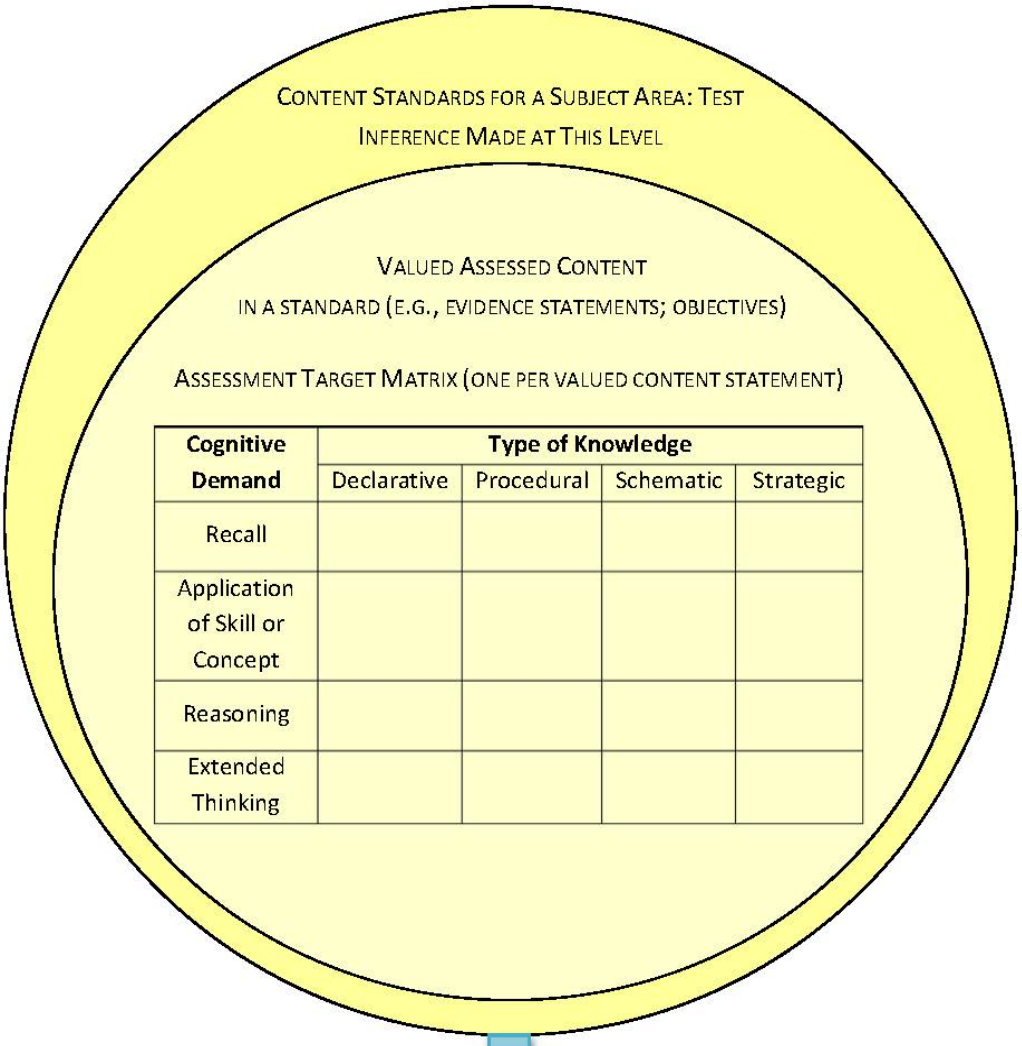
How the framework might be applied can be illustrated using the College Board's (2011) valued science practices for the AP Biology exams. An assessment target can be identified from the appropriate level of the framework as follows. The introduction to the curriculum framework explains, "A practice is a way

to coordinate knowledge and skills in order to accomplish a goal or task. The science practices enable students to establish lines of evidence and use them to develop and refine testable explanations and predictions of natural phenomena (pg 1)." Practice 3 within the curriculum framework, for instance, states "The student can engage in scientific questioning to extend thinking or to guide investigations within the context of the AP course (pg 82)." For each of the valued practices components of the practice are listed, so for practice 3 one of the components is "The student can pose scientific questions (pg 83)."

If this component were to be measured on an AP Biology exam, the College Board might specify the types of evidence that should be evaluated within this component. For instance they may want to see whether students can identify essential features of a given phenomenon. Using the item development framework, test developers would specify where in the assessment target matrix they wanted one or more items designed to produce this evidence at the proper levels of complexity. To produce the items, each assessment target is linked to a set of item writing aspects. The aspects stipulate specific direction to item writers; for example they may include aspects such as use a low degree of abstraction and explicitly request the salient features, but allow for openness to how the features are produced.

Figure 7 illustrates the framework model.

Figure 7: Item Framework Model




| Level | Degree of abstraction | Steps and operations | Solution strategies | Extraneous information | Novelty | Response structure | Response expectations |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | | |
| 2 | | | | | | | |
| 3 | | | | | | | |

## Tier 1

The assessment targets were developed to distinguish facets of complexity so desired evidence can be more explicitly unpacked. We argue that the two dimensions of types of desired knowledge and cognitive demands should anchor the targets. Explained by Ruiz-Primo and others, types of knowledge refer to the kinds of rudimentary to sophisticated knowledge, skills, and abilities (KSAs) we want to have students demonstrate in our tests. Underlying the "types of knowledge" representation are conceptual continua of task or content demands associated with each targeted KSA (for example, see Kennedy & Wilson, 2007; Gotwals & Singer, 2009), but parsing the continua for item writing purposes seems to be too abstract for generalization. The categorization of types of knowledge, on the other hand, is a more user-friendly way of communicating roughly the same concepts, especially as they pertain to test and item development. The second dimension, basic to complex cognitive demands, points to the processes required to engage the various levels of knowledge, skills and abilities. This dimension has been described by several theorists, including Porter (2002); Anderson, Krathwohl, & Bloom, (2005).

We suggest that by crossing these two conceptual continua, a matrix is formed which broadens how complexity is generally explicated in testing circles, and defines range of complexity as products along two related but distinct continua. The target matrix also functions to anchor the work of item writers by both untangling the two concepts and at the same time having them consider and analyze how points along the each of the continua interact with points along the other. More detail about each of these dimensions follows in the Assessment Target section below.

It is our intent that the assessment target matrix is completed as part of the test and/or item specifications development, rather than have the matrix completed by item writers themselves. These targets then become a 'non-negotiable' part of the framework for a particular test in that item writers will be directed to write items that conform to these given cells. In turn, the discrete cells formed in the

matrix will provide more specification to the writers and clarify and constrain the types of evidence expected by the test developers.

## Tier 2

Tier 2 gives concrete direction to item writers. The direction is designed to help writers create the items most likely to produce the desired evidence, and is individualized to the target cell level. The guidance is produced by specifying sets of item feature levels that direct the item writing activity for particular cells. The goal of the direction is to focus and direct the item writing activity, but not overwhelm or confuse it. As such, seven features, explained below, were selected from the literature as those that a) appear to be the most useful and salient for aiding writers, b) are commonly understood when considering how to design item questions, and c) are relatively few in number. For a feature to be included in the framework, it must define a unique aspect of how items function. That is, the item features will most often be correlated, but each must be separable as well.

Each feature falls on a conceptual continuum with two or more discrete levels or markers that operationalize it in such a way that further explains, stipulates and constrains the intent of the feature. The levels generally range from concrete to abstract, or basic to complex. Most often writers will receive details about all features, with delineation of feature levels occurring by cell where the types of knowledge and cognitive demand cross.

These sets of feature levels are called item writing profiles, and are conceptualized such that any one combination of feature levels will not be replicated anywhere else in the assessment target matrix. There are cases where adjusting a level of one feature in a profile would influence the level of another. In these situations an additional profile can be produced that will still effectively provide guidance about item writing within the same cell. Therefore, a cell may contain multiple item writing profiles, but each item writing profile is unique to a specific cell. The intent of the framework is that these compensatory

profiles will all be listed within the cell and item writers can use any one of the profiles in that cell to build a single item.

The framework is designed so that item writers are not free to 'mix and match' among profiles within the same cell. The reason for this is that each profile is balanced to holistically reflect the evidence intended by the test developer. The profile-specific features and their values are meant to operationalize the intent of the target, but it is the cell-specific inference that is driving the decisions about the features. If item writers or writing programs were free to manipulate the item features, levels, and/or their corresponding cell-specific levels, these changes could easily impact the defensibility of the evidence, and consequently the test level inference.

## II. Assessment Targets

As noted above, we have argued that test specification can be improved by identifying the each assessment target for a standard, objective, or essential practice and classifying that target by two dimensions of cognition: type of knowledge and cognitive demand. In designing the test specifications/blueprint test developers identify the assessment targets, categorize these targets by the dimensions in the matrix, and determine for each assessment target how many items should be developed for each populated cell. Depending on the types of and complexity of evidence the test developer is intending, it is not unusual to measure the valued topic or practice with more than one item. As with other aspects of test specification development, these coverage determinations are based on the desired evidence needed to produce viable inferences the assessment user will make from the test scores.

Utility for item development drove the choices about how the two dimensions of cognition are characterized. For instance, the type of knowledge dimension nods to how content or task demands might be conceptualized, and we recognize there are other, more complex, and more theory-deep ways

to explain this continuum. However, we argue that type of knowledge (as defined by Ruiz-Primo and others) is more easily understood and can more easily be explained to and replicated by item writers. Likewise, the cognitive science and learning theory literature have provided a rich resource of theory and empirically -driven information to explain cognitive demand. Nevertheless, for the purposes of the framework, the categorization arising from the assessment literature related to this topic (which we acknowledge is greatly simplified) is more recognizable to educators and test developers, encouraging us to choose an explanation from this literature to describe cognitive demand.

This familiarity of both operationalized dimensions seems to be a major advantage in conceptualizing the Item Development Framework, at least for now.

## Dimension 1: Type of Knowledge

According to Shavelson, Ruiz-Primo, Li, and Avela, 2003; Graf, Peterson, Lawless, Steffen, 2005; Li, Ruiz-Primo, and Shavelson, 2006; Luecht, 2009; Oosterhof et al , 2008; and Ruiz Primo, 2009, items can be characterized by the type of knowledge needed to address them. Four types of knowledge often used to describe task demands are declarative, procedural, schematic, and strategic knowledge. Declarative knowledge is required when tasks focus on having students demonstrate their knowledge of facts, principles, concepts, procedures, and the like. If the goal of a task is to provide evidence of student ability to select and use concepts and procedures, the task will draw on procedural knowledge. Likewise, schematic knowledge is involved in responding to tasks that elicit evidence of student ability to organize and synthesize or connect relevant principles, concepts and/or procedures and apply them to a task. Finally, to elicit evidence of strategic knowledge, tasks focus on having students demonstrate how they solve ill-defined problems, for example by requiring the student to decide how to represent a problem, apply one or more of a number of possible solution strategies, or integrate the other three types of knowledge. Table 1 provides brief descriptions of the four types of knowledge.

Table 1. Types of Knowledge (Ruiz-Primo, 2009, pages 24-25)

| |
|---|
| **Declarative:** *Knowing that*. This type includes knowledge that ranges from discrete and isolated content elements such as terminology, facts, or specific details to a more organized knowledge forms such as statements, definitions, knowledge of classifications, and categories. |
| **Procedural***: Knowing how*. This type involves knowledge of skills, algorithms, techniques, and methods. Usually, it takes the form of if-then production rules or a sequence of steps (e.g., measuring temperature using a thermometer, applying an algorithm to balance chemical equations. add, subtract, multiply, and divide whole numbers). It ranges from motor procedures (e.g., folding a filter paper and placing it in a funnel or classify objects by size), to simple application of a well-practiced algorithm (e.g., subtracting two numbers), to a complex procedure (e.g., implementing a procedure to find out which substance is the most soluble or applying strategies to estimate the results of rational-number computations). Procedural knowledge involves the knowledge of techniques, and methods that are the result of consensus, agreement, or disciplinary norms (e.g., Anderson, L. et al., 2001). It involves how to complete a task and it has been viewed as a skill (e.g., Roger, Ciscero, & Carlo, 1993). Procedural knowledge can be automatized over many trials (practice) allowing retrieval and execution without deliberate attention. Automaticity is considered as one of the key characteristics of expertise (e.g., Anderson, J. R. 1983). |
| **Schematic:** *Knowing why*. This type involves more organized bodies of knowledge, such as schemas, mental models, or "theories" (implicit or explicit) that are used to organize information in an interconnected and systematic manner. This form of organization allows individuals to apply principles or explanatory models to approach a problem (troubleshooting), provide an explanation, or predict an outcome (e.g., explaining why we have seasons; De Kleer & Brown, 1983; Gentner & Stevens, 1983). For example, combined with procedural knowledge, schematic knowledge is involved, in the process of reasoning from several theories to design experiments. |
| **Strategic:** *Knowing when, where, and how to apply knowledge.* "The application of strategic knowledge involves navigating the problem, planning, monitoring, trouble-shooting, and synchronizing other types of knowledge. Typically, strategic knowledge is used when one encounters ill-defined tasks" (Li & Tsai, p. 14). It includes domain-specific strategies, such as ways of representing a problem or strategies to deal with certain types of tasks. It also entails such general monitoring performance or planning strategies as dividing a task into subtasks, reflecting on the process to explore alternative solutions, knowing where to use a particular piece of schematic knowledge, or integrating the three other types of knowledge in an efficient manner. It is important to mention that strategic knowledge, a higher-order knowledge, is based on the other three forms of knowledge (e.g., Anderson, L. et al. 2001; Li, 2001). An attempt to focus only on strategic knowledge without a strong base for the other forms of knowledge does not support transfer to new situations (Mayer, 1997; Pellegrino, 2002). Unlike the other three types of knowledge, "strategic knowledge … is applicable to a wider variety of types of problems within a domain" (de Jong & Ferguson-Hessler, 1996, p. 107). |

## Dimension 2: Cognitive Demand

The assessment literature has a number of models for describing the cognitive complexity of items and tasks, most of which are influenced by Bloom's (1956), hierarchy of cognitive demand (e.g., Anderson, Krathwohl, & Bloom, 2005; Porter, 2002). The model used more frequently in state assessments programs is Webb's (1999) depth of knowledge or DOK model, which defines a hierarchy of demand from Level 1, recall and recognition, through Level 4, extended thinking. According to Webb (2007), depth of knowledge levels reflect an increasing complexity of knowledge required by a task or content standard. Depth of knowledge is not synonymous with difficulty, although the two are related. Hess (2006) describes the relationship between DOK and difficulty as follows[1]:

"The Webb levels do not necessarily indicate degree of 'difficulty' in that Level 1 can ask students to recall or restate a much more complex concept, the latter being much more difficult. Conversely, in depth of understanding, a concept is required to be able to explain how/why a concept works (Level 2), apply it to real-world phenomena with justification/supporting evidence (Level 3), or to integrate one concept with other concepts or other perspectives (Level 4)" (page 4, italics in original).

One advantage of using the Webb model is that a majority of state departments of education have used it as part of evaluating the alignment of assessments to standards – the model, or its variations, are familiar. Webb (2002b) has defined the DOK levels a bit differently for different content areas (e.g., mathematics, writing). However, the intended purpose of thinking required in each level across content areas is the same, and we have fashioned a slightly modified hierarchy using Webb's general definitions (Webb, 2002a) and his more specific, content-related definitions (Webb, 2002b), filling in gaps or further

[1] The levels described by Hess correspond to the Levels 1-4 listed in Table 2, although the language in the two descriptions is not identical

defining levels using related concepts from other sources (for example, Jonassen & Hung, 2008; Oosterhof et al, 2008; Anderson et al., 2005; Hess, 2006).

Table 2. Cognitive Demand (adapted from Webb, 2002b)

| | |
|---|---|
| Level 1: Recall | Recall or reproduction of information – e.g., facts, definitions, concepts, simple procedures—and performing simple processes that are generally single-step, close to automatic or habitual (e.g., simple formula, reading aloud w/o comprehension required). |
| Level 2: (Application of) Skill/Concept | Beyond recall or reproduction, this level requires a decision to be made regarding how to approach the problem. It also requires both comprehension and processing, for instance, interpreting relatively uncomplicated information, giving examples of a concept, state and explain relationships, collecting and organizing data, and comparing surface features of concepts or ideas. |
| Level 3: Reasoning | Items at this level of cognitive demand require reasoning, planning, and using evidence, complex and abstract thinking. Tasks include justifying a response, explaining concepts in terms of phenomena, generalizing across and connecting ideas, and applying concepts in new situations. |
| Level 4: Extended Thinking | Extended thinking requires complex reasoning, higher order thinking, typically making a number of connections and devising a solution approach from among many alternatives. Typically, these tasks take an extended period of time. Examples include designing and conducting an experiment, analyzing and synthesizing information from a number of sources, and applying information from various sources to develop a new concept. |

## The Assessment Target Matrix

The crossing of the two dimensions just summarized results in a matrix that is used to characterize and unpack the complexity of the assessment targets (see Table 4). Where one or more assessment targets related to a particular topic falls in the matrix depends on the *primary* intent of each target. This means that more than one type of knowledge may be implied by an assessment target, especially when higher types of knowledge are intended. This is because the application of higher types of knowledge often requires the use of lower levels of knowledge, for instance, applying schematic knowledge to a problem almost always includes the application of declarative and procedural knowledge.

Further, because the type of knowledge and cognitive demand dimensions are related, there are some combinations of the two dimensions seem to be unlikely. The highlighted cells in Table 3 suggest a stepwise relationship that indicates where most assessment targets will logically fall. It is not yet clear if assessment targets might reflect the Strategic/ Application of Skills and Concepts cell.

Table 3. Assessment Target Classification: Targeted Types of Knowledge and Cognitive Demand

| Cognitive Demand | Type of Knowledge | | | |
|---|---|---|---|---|
| | Declarative | Procedural | Schematic | Strategic |
| Recall | | | | |
| (Application of) Skill/Concept | | | | ? |
| Reasoning | | | | |
| Extended Thinking | | | | |

## III. Direction to Item Writers

The second essential component of the item development framework is the differentiated item writing direction supplied to writers. The distinctions in the specifications are made at the assessment target matrix cell level, and the guidance to writers will be in the form of item writing profiles that specify direction along seven item feature continua. The features outlined below have been identified as among the most salient aspects for item writing, with a level of detail and clarity that is simultaneously comprehensive enough to cover critical aspects of items and tasks and constrained enough to be feasible and usable for item writers. For this guidance to be successful and effective, it ultimately needs to produce items where the evidence is consistent with and aligned to the intent of the cell in which the assessment target resides, and where the evidence that contributes as expected to the test level claims and inferences.

As the literature review illustrates, the learning, cognition, and assessment literature is replete with explanations of how one might characterize aspects of knowing and skill development, and sometimes how one might elicit evidence of certain types of knowledge, skills, and abilities from students. Depending on the goals of the work, the grain size of the information ranges from very nuanced to broadly expressed. For the most part, we found finely attuned descriptions of item features to provide richness and depth, and general descriptions to provide a sense of subject, usually oversimplified. Mid-level categorizations seemed most useful to us, when they made connections to learned academic content, concepts and abilities, and to the form of interactions between questioners and responders. The work of authors such as Shavelson and others (2003) Oosterhof et al. (2008, 2011) , and Luecht and others (e.g. 2005, 2006, 2007, 2009) seemed to exhibit more mid-level characterizations of features, whereas descriptions of detailed aspects came more from researchers such as Feltovich et al (1993), and Ederbach and Crowley (2009).

## Specification of Item Features and their Levels

The explanation of the features and the continua they reflect are a first draft; they represent characteristics of items that are likely to affect the type of knowledge and cognitive demand required by the item. Table 4 presents the list of item features and their levels as we have conceptualized them to date. These were developed based on our reviews of the literature and iterative refinements after internal tryouts with available items. The features affect the items in several areas of student/item interaction: understanding the nature and requirements of the problem, or problem apprehension; devising and carrying out a solution strategy for the problem; and providing a response to the problem. It is envisioned that the specifications will be accompanied by supplementary materials explaining in more detail what each feature continuum reflects. In many cases, the features are amalgamations of related qualities, differentially used and expressed, and examples of these qualities will be part of the materials.

Table 4: Item Features and Levels

A. Degree of abstraction in the presentation of the subject matter and nature of the task to be addressed
   1) The information presented is primarily concrete.
   2) A mix of concrete and abstract information is presented.
   3) The information is presented primarily abstractly.

B. Characteristics of the steps or operations typically required to solve the problem
   1) One
   2) Multiple sequential
   3) Simultaneous

C. Expectations for solution strategies typically used to address the task (Number of ways to successfully complete the task)
   1) One or two strategies are typically used to address the task.
   2) There are a limited number of strategies and/or combinations of strategies that can be address the task.
   3) There are numerous strategies and combinations of strategies that can be used to address the task.

D. Use of intentional extraneous information
   1) No
   2) Yes

E. Novelty of context, content, or application approach
   1) The context and/or application are likely to be very similar or the same as those used in the classroom.
   2) Some aspects of the context and/or application are likely to be very similar or the same as those used in the classroom.
   3) The context and/or application are likely to be novel to the student.

F. Level of detail about what the outcome response should consist of
   1) Specific details about what the outcome response should consist of are stated directly.
   2) Some inference is required to determine what the outcome response should be; many of the details are specified so the level of inference needed is not terribly high.
   3) Inference is required to determine the details of the task's goal(s); the goal is stated broadly.

G. Explicit and specific response expectations
   1) The task directly states specific actions to be performed.
   2) Some inference is required to determine what actions to take to address the task; the task states most procedures directly or the level of inference needed is low.
   3) Inference is required to determine what actions to take to appropriately address the task.

*The first feature*, degree of abstraction is discussed by several authors including Feltovitch et al (1993), Famularo & Russell (January 2007), Marshall, (1993), Pellegrino et al (2001), and Lee et al (in press). Luecht (2009) also refers to information density that reflects conceptual richness in terms of relations and variable elements. This feature intends to encompass several different aspects of structure of knowledge, including discreteness to continuity, types and complexity of relationships between observed and implied knowledge, skills, and ability elements, and number of relevant ideas and links required among the ideas. As we conceptualized the feature, we decided that, while there were many ways to depict level of abstraction, choosing among those ways was the task of the item writer, not this framework, since degree of abstraction is at least partially dependent on the content being measured.

*The second feature* focuses on the cognitive or procedural manipulations typically required to arrive at the solution. Other item features affect the details of the steps and operations required. Clearly, the nature of the steps is dependent on the topic, type of knowledge, and cognitive demand, such that recalling or applying steps in a learned mathematics operation is different from reasoning from results compiled based on carrying out a plan that assembles data from divergent sources. This feature focuses on what the student needs to do given what is presented to him/her in the item. Many of the same authors referenced above discuss this aspect as well, as do Eberbach and Crowley (2009) and Oosterhof et al (2008). Other than the one or two-step operation, the continuum of complexity as reflected by the literature does not seem to be based on number of steps so much as it is on how the steps unfold, that is, sequentially or simultaneously.

*The third feature*, expectations for solution strategies used to address the task, refers to the openness of the task and the degree to which a student needs to identify a strategy as opposed to being directed in the item or being constrained by the narrow nature of the topic or task. Shavelson and colleagues (e.g. Stecher et al, 2000 Li, Ruiz-Primo & Shavelson, 2006; and Ruiz-Primo, 2009) and Cavagnetto (2010) in

particular discuss this nature of degree of support. Like the first feature, at a given level of support, the topic content or approach the item uses to communicate the problem isn't as central here as the range of specificity regarding the strategies students may be asked to utilize.

*The fourth feature*, use of extraneous information, is a technique used to measure students' ability to recognize and evaluate what is salient information that is necessary for the purposes of a particular item. This feature can add complexity by introducing context that may or may not be central for understanding the topic or task but in any case is not central for producing a solution. Complexity can also be impacted by introducing variables, connections or relationships that function as distractors. Correctly incorporated, extraneous information can add richness to the context or surround within which related items sit, or it can add unnecessary confusion if it is not used intentionally or judiciously. This feature, as are all the features, reflect aspects that are intended to affect what is measured in a construct-relevant way, Use of extraneous information is discussed by several authors, including Luecht (2009), Ederbach and Crowley (2009), Oosterhof (2008) as well as by Shavelson and others. Unlike several of the other features though, use of extraneous information is represented as dichotomous -- simple presence or absence.

*The fifth feature*, novelty of context, content, or/or application approach, is also relatively common in the literature we surveyed. Ferrara and Duncan 2011), Snow et al (2010), Cavagnetto, (2010), among others discuss degree of novelty in terms of measuring deep mastery of the underlying principles of a topic, concept, or skill by properly applying them in unfamiliar settings or for unfamiliar uses. Like extraneous information, novelty is a technique meant to impact certain aspects of complexity. Further, these two features appear to help confirm or otherwise nuance and temper earlier features of abstraction, steps or operations, and solution strategies.

*Features six and seven* refer to the response parts of items. Feature six relates to the degree to which the item is explicit about the nature and form of a correct response. In contrast, feature seven refers to the degree to which the item is explicit about the actions that need to be taken to arrive at the correct response.

Each of these response features are related to earlier features, in particular feature three, expectations for solution strategies. Gitomer and Rock (1993), Shavelson and colleagues, Oosterhof et al. and Snow et al (2010), among others, discuss the intent of the sixth feature and how the nature of the continuum from explicit to implicit constrains or opens up the degree to which a student needs to conceptualize, organize and implement solutions based on the overall intent of the item. Feature three refers to degree of openness when it comes to how the task presented in the item constrains problem solving, whereas feature six refers to openness of how to conceptualize the correct solution.

Many of the same authors discuss feature seven as well, where the nature of the actions to arrive at the correct response are the focus. Again, while feature three focuses on the breadth of strategies that are possible, given the particular task demand, feature seven focuses on how explicit is the item about what types of strategies to use, regardless of whether the appropriate strategies are few or many.

## Features and the Response Space of Items

The last two features, in particular, give a nod to direction about response considerations, but we are aware that this is inadequate guidance with respect to how to choose and design a response environment that is consistent with the intent of items shaped by the features and their levels. As part of our literature review, we did search for ways authors suggested how the response environment might be characterized in items, and used this information to jump-start a description of response space features. However, other than suggesting general item genre types, we didn't see much more. We then considered whether providing some type of suggestions about item types for each cell may be

sufficient—not only specifying general genres (for instance, multiple choice or constructed response), but suggesting more specific attributes of genre types such as multiple response multiple choice, or extended constructed response. Nonetheless, we remain concerned that this may not be specific enough direction, as we know that using the wrong type of response space can distort even the most well-conceived explanation of the problem, stem, or prompt.

For this project we were not able go any further. Yet, we did extend the question about how to characterize the response environments of items being reviewed in our initial proof of concept review relative to how the reviewers evaluated the intent of the writers as per the features identified above. Our questions to the reviewers can be found in Table 5. We reasoned that comments from the reviewers may help guide the next stage of development of the framework and the features in particular, providing us with some sense of how to focus additional response space features, as part of next steps.

Table 5: Review Questions About the Response Environment

| Response Considerations<br>Please characterize the response space of each task/item you review. Sample considerations are outlined below. |
|---|
| • How consistent is the response environment (close-ended formats with key/distractors, spatial constraints for constructed response formats, or other response elements used in open response TEs) with students being able to satisfactorily fulfill expectations in 'B'? Please explain.<br><br>• How consistent is the response environment (close-ended formats with key/distractors, spatial constraints for constructed response formats, or other response elements used in open response TEs) with students being able to satisfactorily fulfill expectations in 'C'? Please explain.<br><br>• How consistent are the response key/distractors for close-ended formats or rubrics for constructed response formats with directions given in 'F'? Please explain.<br><br>• How consistent are the response key/distractors for close ended formats or rubrics for constructed response formats with directions given in 'G'? Please explain. |

Although the literature from which these features were culled was not meant to be exhaustive, the insights the review represents seem to reflect a reasonable range of consensus from which to begin. As such, we argue that the identified features represent a core of characteristics that, taken as a whole, include a number, if not enough or all, of the most significant signposts that might underpin more principled item writing. At the same time, the features and the continua they embody seem rich enough to provide differentiated direction to writers by intent of each cell, at least intent in terms of cognitive demand and types of knowledge. Clearly, the next steps, besides revising existing features based on the proof-of-concept review and rounding out the features related to the response space, are to continue the tryouts with the features and levels using existing items, elicit reviews and comments from experts, add specificity and exemplars around the features to prepare for training item writers, and subsequently evaluate their utility.

## III. External Proof-of-Concept Review

The purpose of the external review of the framework was to evaluate if the ideas and materials shared here appear to be clear and accessible to item writers, and what still needs substantial work. The review was undertaken by asking writers to not only evaluate the concepts in the framework, but also to use the framework to evaluate some existing items. While evaluation of items is not the ultimate intent of the framework, we used this exercise to prompt more detailed comments from our reviewers about the usability of the method and its various elements.

For this activity, 26 existing items from mathematics and science at elementary, middle school, and high school levels were purposely selected to cover a range of matrix cells. Three writers were identified but time commitments precluded one writer from participating. Each writer has expertise in developing mathematics and science items, and the existing items were familiar to the reviewers. To train the writers, five documents were compiled and can be found in Appendices I.3-I.6:

1) An overview of the framework
2) The assessment target matrix, a short explanation of the purpose of the assessment target matrix, each dimension, and a sense about how the intersection of the two dimensions might be read
3) The list of item features and their levels, a short explanation of the purpose and meaning of the features and levels
4) Five exemplars of existing items with internal subgroup evaluation results by cell, features, and levels to be used for training
5) A tryout chart to be completed for each item the writers evaluate. The chart contains places for feature and response space notes.

The writers met with one of the PIs in summer 2012 and were trained using these materials. Training took approximately two and a half hours and consisted of review of the materials and completing an evaluation of three of the exemplars items together. At that time the writers individually spent time reviewing the materials and completing the rest of the exemplars on their own. Subsequently, the reviewers and trainer reconvened to clarify and answer questions. The writers were then asked to begin work with the 26 items; the trainer met with the reviewers after a week to address any questions. The writers work in the same office, and while they were asked to do their own work, they were told they could clarify the meanings of the matrix dimensions and cells and meanings of the features and levels with each other and/or the trainer. By the end of the training and follow-ups, the trainer and the writers felt that the writers understood the material sufficiently. However, there were no formal calibration reviews to determine if the writers had drifted from this understanding prior to completing their task.

## Preliminary Findings

The evaluations from the reviewers can be found in Appendix I.7. Placement of items in the assessment target matrix cells found 100% agreement. This is not surprising given the writers' familiarity with the type of knowledge and cognitive demand concepts, familiarity of the items, and access to the other writer to build consensus of how to interpret the matrix. Of interest, however, was where the items fell on the matrix relative to how the subgroup anticipated that items would typically populate it. The differences can be found in Table 6. The subgroup hypothesized that the greyed cells would be where most if not all items would fall; the numbers indicate are where the writers placed the 26 items. In three of the four cells subgroup members considered to be improbable, writers found items they believed fit the cell intent.

Table 6. Tryout Findings of Assessment Target Classification of 26 Items

| Cognitive Demand | Type of Knowledge | | | |
|---|---|---|---|---|
| | Declarative | Procedural | Schematic | Strategic |
| Recall | 2 | 2 | 2 | |
| (Application of) Skill/Concept | 3 | 4 | 5 | ? 1 |
| Reasoning | | 3 | 4 | |
| Extended Thinking | | | | |

The item evaluations and notes from the reviewers indicated much more variability between the results than was the case for the assessment targets. There were a few places both reviewers considered either problematic or sufficient, however. First, reviewers were confused about the last feature, explicit and specific response expectations. The writers each chose a different level in 22 of 26 items, with non-adjacent levels in 3 cases. In their debrief the writers said the general concept of the feature was viable, but that they got confused by how it was to be applied and the choices they had to select from. Conversely, they found both the task and levels for the third feature, expectations for solution strategies typically used to address the task, to be clear and well-grounded. They also felt that the extraneous information and novelty features (features 4 and 5 respectively) were relatively straightforward; it seemed that most differences in level choices between the reviewers could be avoided by having a source they could ask about if the item information and situations were typical or not.

For the second feature, characteristics of the steps or operations typically required to solve the problem, they raised the same point as a couple of the subgroup members had—that is, they would have preferred if the first level was one or two steps, with the second level multiple sequential and greater than two steps. They commented that they felt they needed better training on both the first feature that asks about degree of abstraction, and the sixth feature that asks about level of detail about an

outcome response. Specifically, they did not have enough understanding of what constitutes 'abstraction', and 'detail', and they said that the levels in the sixth feature were not well stated. One of the writers felt more comfortable about the differences between the sixth and seventh response features. One point the trainer heard again and again is that the differences between levels was often a judgment call, and in many cases they could have made an argument for each of two adjacent levels depending on how they perceived the intent of the feature. Further, they agreed that formal calibration during the evaluation process would have been helpful.

The writers were asked about how they might find and use the matrix, features and levels if they were presented with them as direction for writing items. Both reviewers had a hard time answering this question. They found the features interesting, but weren't sure they would be adequate for this charge. They didn't want to see levels of the 6 features reduced to two, but weren't sure if only three were 'good enough' for writing purposes. They did find the matrix useful for parsing out what complexity means and helpful for understanding what a test developer would want to see in items in particular cells. Neither writer made written comments about the response considerations relative to the intent of the items. This could be because of their familiarity with the items and that they were trained to consider the choice of response space as they were building items. In the debrief, they agreed that aligning the response space to the intent of the stem or explanation of the problem was essential, but they didn't have any ideas for how to put these considerations into features.

## V. Closing Comments

In this chapter we attempted to identify and explain our proposed framework for item development. We began this work because we saw a gap in the literature. Authors in several related fields associated with learning have presented and explained in rich detail the cognitive and pedagogical aspects of how students acquire knowledge, skills and abilities, integrate more specific and broader ways of thinking

associated with these KSAs, and conceptualize and nurture the development of mastery in particularly science and mathematics. Further, a number of authors from the assessment field have recently worked to produce a series of first rate item evaluation systems aimed at more fully understanding that how we ask students questions matter, both in terms of which item characteristics affect what students understand about an item and how characteristics seem to interact with one another to produce effective, and ineffective, items based on particular intent of what is to be measured.

Both of these endeavors have been essential in helping us to appreciate how students learn, know and develop skills in academic settings, and have laid the groundwork for how we might understand what they know. What appears to be largely missing to date, however, is producing more specific, literature-based, usable, and effective guidance for writers of test items. It is one thing to analyze the nuances of what makes existing items work or not; it is quite another to produce a principled approach that is effective in giving sound, specific direction to the item writing activity, while still being user-friendly and understandable. If properly conceptualized and operationalized, we argue that producing such a project can give the assessment field and its consumers greater confidence that the evidence resulting from items written or generated under such an approach can be more directly linked to the types of content and cognitive demands we are intending to measure in academic assessments.

We contend that, for a number of reasons, the Item Development Framework introduced here is a step in this direction. First, academic complexity, as a function of less or more challenging content and something associated with item difficulty, has been mostly conceptualized along one index or another. It seems that crossing task demands (by proxy, type of knowledge) and cognitive demand is a valuable way of nuancing and parsing out aspects of complexity from which to write items.

Second, the standards, specifications, and assessment target tier and the item writing specifications tier together provide substantially more information about targeted claims than do typical test development

procedures. The first tier, as noted above, requires that test developers consider not only the topics by the standards they want to sample, but then also specify the particular complexity cells of the assessment matrix they want evidence within, and therefore items written to. What is important here isn't that the test level specifications are unpacked by standard, topic, and complexity. In and of itself, drilling down in this way isn't unique. What is novel is a) the unpacking of complexity; b) explicitly linking tier one information at the topics level with the test level inferences above it; and c) differentially adjusting the item writing specifications of the tier 2 as per the cell placement of the in the first tier.

Third, the intent of the project is that the distinctions and qualities of the model overall and within each tier leads to more differentiated and detailed direction to item writers as opposed to how item writers are typically trained and guided now. The framework gives writers more specific information about how to translate test specifications, how to write to particular cell specifications, and ideas of where compensatory tradeoffs of item traits are by comparing profiles in cells. We argue that these steps reduce the opportunities for misunderstandings within and across writers, and within and across topics and standards.

Clearly, the proof-of-concept results tell us we still have work to do in clarifying and more fully explaining parts of the framework. The writers also were silent on how to include additional features to help writers align response spaces to what they are intending the items to measure. Having at least three levels per feature seems to be a viable approach, but more refinement is needed in writing the levels, and more training time, including use of systematic calibration sets, seems to be called for until the concepts embodied in the features become more commonplace in test writing circles. The reviewers could not tell us if the features were good enough for purposes of writing items.

Recent test specifications of academic assessment consortia and other agencies involved in achievement testing, for instance the College Board, seem to be moving in this direction. By stipulating more

specifically the types of evidence that are required for different topics within particular complexity levels either explicitly or implicitly defined, these details provide more information for item writers to complete their work. While the direction to item writers is neither as direct nor detailed as we are proposing, the intent of improving how we collect better evidence, and evidence we can trace back to test level inference, seems consistent. This is good news.

We have anticipated several next steps. Based on our assessment of the results from this project, we suggest that future iterative work on the framework will refine the features and their levels, tentative profiles will be expanded and refined, and evaluation tryouts using more items will occur with further refinement of all framework elements. Somewhat concurrently, we expect that material and exemplars for item writing training will be developed and piloted, and then item writers will test out the item framework and their work will be evaluated for effectiveness, thoroughness, utility and feasibility.

# References


Alves, C. B., Gierl, M.J., & Lai, H. (2010). Using Automated Item Generation to Promote Principled Test Design and Development. Paper presented at the annual meeting of the American Educational Research Association Denver, CO, April 30 – May 4, 2010.

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2005). *A taxonomy for learning, teaching, and assessing*. Longman.

Baxter, GP & Glaser, R (1997). An approach to analyzing the cognitive complexity of science performance assessments. CSE Technical Report.

Bloom, B. S. (1956). Taxonomy of educational objectives: the classification of educational goals (Vol. 1). Longmans, Green.

College Board (2011). AP Biology, Curriculum Framework 2012-2013. College Board: Washington, D.C.

Li, M., Ruiz-Primo, M.A., & Shavelson, R.J. (2006). Towards a science achievement framework: The case of TIMSS 1999. In S.J. Howie and T. Plomp (Eds.), *Contexts of learning mathematics and science: Lessons learned from TIMSS*. Florence, KY: Routledge.

Famularo, L., & Russell, M. (January 2007) Examining the Utility of a Prototype Assessment for Assessing Students in the Gap. Report of the Technology and Assessment Study Collaborative, Boston College.

Feltovich, P.J., Spiro, R.J., & Coulson, R.L. (1993). Learning, teaching, and testing for complex conceptual understanding. In Frederikson, N., Mislevy, R.J., & Bejar, I.I. (eds.). Test Theory for a New Generation of Tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ferrara, S., Duncan, T.G., Freed, R., Vélez-Paschke, A., McGivern, J., Mushlin, S., Mattessich, A., Rogers, A., & Westphalen, K. (2004). Examining Test Score Validity by Examining Item Construct Validity: Preliminary Analysis of Evidence of the Alignment of Targeted and Observed Content, Skills, and Cognitive Processes in a Middle School Science Assessment Paper presented at the annual meeting of the American Educational Research Association, April 2004, San Diego, CA

Ferrara, S. & Duncan, T. (2011). Comparing the science achievement constructs targeted and achieved in a statewide middle school science assessment. *Education Forum*.

Ferrara, S., Huff, K., & Lopez, E. (2010). Targeting Cognition in Item Design to Enhance Valid Interpretations of Test Performances: A Case Study and Some Speculations. Paper presented at the annual meeting of the American Educational Research Association, Denver.

Frederikson, N., Mislevy, R.J., & Bejar, I.I. (1993). Test Theory for a New Generation of Tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jonassen, D.H. & Hung, W. (2008). All problems are not equal: Implications for problem-based learning. The Interdisciplinary Journal of Problem-based Learning, *2*, 6-28.

Kennedy, C.A., & Wilson, M. (2007). Using Progress Variables to Interpret Student Achievement and Progress. BEAR Technical Report No. 2006-12-01. University of California, Berkeley.

Li, M., Ruiz-Primo, M.A., & Shavelson, R.J. (2006). Towards a science achievement framework: The case of TIMSS 1999. In S.J. Howie and T Plomp (Eds.), *Contexts of learning mathematics and science: Lessons learned from TIMSS*. Florence, KY: Routledge.

Li, M., and Tsai, S. (2007, work in progress). Linking Assessment to Science Achievement: A Knowledge-Based Approach. Project Report Submitted to NSF.

Luecht, R.M. (2009). Adaptive Computer-Based Tasks Under an Assessment Engineering Paradigm. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Mislevy, R.J. & Haertel, G, D. (2006). Implications of Evidence-Centered Design for Educational Testing. Educational measurement: Issues and Practices. Winter 2006, 6-10.

Mislevy, R., Steinberg, L., Almond, R. G., Haertel, G. D., & Penuel, R. (2003). *Leverage points for improving educational assessment (PADI Technical Report 2).* Menlo Park, CA: SRI International.

National Research Council. (2006). *Systems for State Science Assessment.* Committee on Test Design for K–12 Science Achievement. M.R. Wilson and M.W. Bertenthal, eds. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. Pellegrino, J., Chudowsky, N., and Glaser, R., eds. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2011). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Oosterhof, A., Rohani, F., Sanfilippo, C., Stillwell, P., and Hawkins, K. (2008). The capabilities-complexity model. Florida State University: Center for Advancement of Learning and Assessment.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher* 31, 3-14.

Ruiz-Primo, M.A. (2009). Towards a framework for assessing 21st century science skills. Paper presented for The National Academies.

Seeratan, K., & Mislevy, R. (2009). Design Patterns for Assessing Internal Knowledge Representations (PADI Technical Report 22). Menlo Park, CA: SRI International.

Shavelson, R.J., Ruiz-Primo, M.A., Li, M., & Ayala, C.C. (2003) Evaluating new approaches to assessing learning. CRESST/CSE Report 604. Los Angeles, CA: The Regents of the University of California.

Stecher, B.M., Klein, S.P., Solano-Flores, G., McCaffrey, D., Robys, A., Shavelson, R.J., & Haertel E. (2000). The Effects of Content, Format, and Inquiry Level on Science Performance Assessment Scores. Applied Measurement in Education, *13*(2), 139–160

Webb, N. L. (2002a, April). An analysis of the alignment between mathematics standards and assessments for three states. In *annual meeting of the American Educational Research Association, New Orleans, LA*.

Webb, N. L. (2002b). Depth-of-Knowledge Levels for Four Content Areas. Unpublished manuscript.