APPLIED MEASUREMENT IN EDUCATION, 23: 325–341, 2010

ISSN: 0895-7347 print / 1532-4818 online
DOI: 10.1080/08957347.2010.510959

Routledge
Taylor & Francis Group

Representing Targets of Measurement Within Evidence-Centered Design

Maureen Ewing, Sheryl Packman, Cynthia Hamen and Allison Clark Thurber
The College Board

In the last few years, the Advanced Placement (AP) Program® has used evidence-centered assessment design (ECD) to articulate the knowledge, skills, and abilities to be taught in the course and measured on the summative exam for four science courses, three history courses, and six world language courses; its application to calculus and English language arts subjects is in progress. The purpose of this article is to describe the methodology that was used with subject-matter experts (SMEs) to articulate the content and skills important in the domain, and then the iterative processes that were used to articulate the claims and evidence to represent the targets of instruction for AP courses, and by extension, the targets of measurement for the AP exams. Discussion will focus on how the use of ECD provides a strong foundation for ensuring the alignment among curriculum, instruction, and assessment while at the same time enhances the validity argument for test score interpretation.

INTRODUCTION

Each step in the test development process is governed by key decisions that are made early on about test design: What is the purpose of the test? What are the intended uses for resulting test scores? What are the content and skills (i.e., the domain) to be measured by the test? Traditional approaches for determining the content and skills to be measured by an achievement test, or the "targets of measurements" for a given domain, include gathering input from subject-matter experts (SMEs) who serve on test development committees, reviewing textbooks

Correspondence should be addressed to Maureen Ewing, The College Board, 661 Penn Street, Suite B, Newtown, PA 18940. E-mail: mewing@ collegeboard.org

and other curriculum documents (e.g., course syllabi, state and national standards), and using empirical results from large-scale curriculum studies (Schmeiser & Welch, 2006).

Regardless of the methods or combination of methods used to identify the targets of measurement, the result is typically the same and involves a listing of content topics, skills, or both that should be measured by the test. Once this information is gathered, the test design phase is largely complete and the test implementation phase begins whereby developers use this information, along with other information about statistical and item specifications, to write test items. However, a simple listing of content topics and skills to represent the targets of measurement lacks the specificity and transparency that can be achieved by adopting evidence-centered assessment design (ECD) (Mislevy & Riconscente, 2005; Steinberg et al., 2003). For example, use of ECD forces answers to questions like: How should content and skills be integrated? What does a student need to do or say to demonstrate knowledge acquisition?

ECD relies on the specification of an evidentiary argument that links features of the task (e.g., a test item) to evidence required to support claims about what students know and can do in the domain. More specifically, the evidentiary argument includes three components: (a) the claims we want to make about students' knowledge and skill proficiencies; (b) the observable evidence within student work required to support the claims we want to make; and (c) the tasks or situations designed to elicit the required evidence. For assessment development, this information is extremely useful because it goes further than just listing content and skills by explicating, via the claims, the interaction between content and skills (i.e., how students reason about the content). The meaning of test scores is clearer and more useful because the evidentiary argument is based not on a general claim that the student "knows the content," but on a comprehensive and useful set of specific claims that indicate specifically what the student knows about the domain. Furthermore, the characteristics of how students can demonstrate acquisition of the content and skills are so clearly identified by the evidence that the resulting documentation is transparent, and item, scoring, and rubric development is strengthened.

The claims that are created through the ECD activities are guided by the purpose (e.g., assessing progress or end-of-course knowledge) and audience of the test (e.g., students, teachers, college admissions officers). For example, claim development can be focused at the summative level or formative level, and can be inputs to both assessment and curriculum design. A focus on the former requires writing claims that represent all of the content and skills that should be acquired after learning a particular domain or the summative goals for a course, while a focus on the latter requires writing claims that represent subsets of the content and skills constituting the formative goals for a course. Summative claims for a final course assessment can be decomposed into formative claims to

guide the development of instruction and assignments that will prepare examinees over a period of time for the summative assessment. In the case of formative claims, the purpose is to guide teaching and assess progress primarily to provide feedback to students and teachers; whereas for summative claims, the purpose is to express summative proficiency in terms meaningful to a variety of audiences (e.g., students, teachers, parents, college administrators).

In the past few years, the AP Program has leveraged ECD to document the content and skills to be taught in AP courses and measured on the exams. The AP Program is unlike most other large testing programs because in addition to determining targets of measurements for its tests it must also produce curricular requirements and course descriptions so that high school teachers can effectively teach the course. As a result, it is necessary for design work to lay a foundation for both course and exam development. AP courses and exams offer high school students the opportunity to take college-level courses while still in high school as well as the opportunity to demonstrate proficiency in the domain by performing well on the corresponding end-of-course exam. Currently, there are over 30 AP courses and exams, and, to date, ECD has been applied to 13 AP courses and exams including four science courses, three history courses, and six world language courses. Similar work with calculus and English language arts subjects has also begun.

Although conceptually the fundamental principle underlying ECD is straightforward, that is, that assessment is a form of making inferences from imperfect evidence, implementation is resource-intensive, iterative, and challenging. It requires a steep learning curve as SMEs must embrace new terminology and methods. The purpose of this article is to describe the methodology that was used with SMEs to articulate the content and skills[1] deemed important in the domain. Next, the iterative processes used by these SMEs to articulate the claims and evidence to represent the targets of instruction for AP courses, and by extension, the targets of measurement for the AP exams, are explained. In other words, the claims and evidence developed to represent the targets of instruction are intended to be the same as the targets of measurement; however, the design process does allow for the possibility that some targets of instruction may only be assessable by classroom assessments and projects and not by AP exams given the constraints of large-scale, national assessments. This topic is further addressed by Huff, Steinberg, and Matts (2010/this issue) and Hendrickson, Huff, and Luecht (2010/this issue).

[1]The word "skills" in this article is used in a general way to refer to the set of reasoning processes or practices that are important for students to be able to use or apply when interacting with content. In science, these skills were referred to as scientific practices; whereas in history they were called historical thinking skills.

Several scholarly papers provide thorough overviews of ECD (e.g., Mislevy, Almond, & Lukas, 2004) as well as examples of using ECD to construct tests (e.g., Steinberg et al., 2003). This article focuses on the first two ECD activities: domain analysis and domain modeling. The products of these activities are called "artifacts." For the domain analysis, the artifacts are a prioritization of content and skills, and for the domain model, the artifacts are the claims, evidence, and achievement-level descriptors (the development of the latter is discussed in Plake, Huff, & Reshetar, 2010/this issue]). This article will provide a comprehensive description of the processes that can be used to write claims and evidence, as well as how these processes vary according to the nature of the domain and the current status of the subject-specific AP course and exam. For example, although the end result was similar, the processes differed to some degree based on whether the AP course already emphasized skills (e.g., world language, Spanish Literature, English language arts) or whether the AP course sought to update courses by making the skills a more prominent part of the course (e.g., history and science). To illustrate the different processes that were used, specific examples are infused throughout the article from various subject areas. Special focus will be given to describing the directions that were given to SMEs to guide their work. The discussion will focus on how the use of ECD provides a strong foundation for ensuring the alignment among curriculum, instruction, and assessment while at the same time enhancing the validity argument for test score interpretation.

METHODOLOGY

Subject Matter Experts

The first step was to convene SMEs for each subject. Each group of SMEs represented both college faculty and high school AP teachers who were recruited based on their reputation and experience in the discipline. The groups were comprised of a diverse set of 4–5 AP high school teachers and 4–5 post-secondary instructors representing different school types and geographic regions. Obtaining perspectives from both college faculty and high school teachers was essential, particularly because high school teachers work directly with the target students and deliver the course content while college faculty teach the courses that the AP Program seeks to represent in the design of their courses. It is important to emphasize the panels of SMEs were not test development committees. In fact, trained item writers were not involved at this stage of the work because the focus of the work was on articulating the domain for the AP course and exams. The role for item writers came later when the work shifted to writing tasks that elicited the evidence to support the claims (see Hendrickson et al., 2010/this issue).

Domain Analysis

The main goals of the domain analysis were to outline the content that is learned in rigorous, entry-level college courses within each subject, while prioritizing depth of understanding over breadth of content, and to identify the skills that are developed within the course or the way in which students interact with the content. Along with these general goals, a specific charge was developed for each group of SMEs to advise on a variety of teaching and learning issues such as "What understandings of the products, practices, and perspectives of the culture studied would students be expected to demonstrate in Spanish literature?" or "How are the essential components of scientific inquiring and reasoning applied within chemistry?" For courses like science and history, a critical goal that guided much of the domain analysis was developing a domain that moved away from content facts and toward one that emphasized students' ability to reason with and apply content in various ways.

Once convened, the first task for the SMEs was to analyze the domain, which involved identifying the content and skills that should be acquired in the AP course and that represented best practices in teaching and learning in the discipline. A report from the National Research Council's Committee on Foundations of Assessment (National Research Council, 2001) operationalized best practices by specifying that deep conceptual understanding is facilitated when instructional methods organize learning around "big picture" ideas and take into account the different ways in which students learn.

Conducting the domain analysis involved gathering information from a variety of sources. The current AP course description, national standards, College Board standards,[2] and the latest research on student learning and assessment were consulted. For example, the American Council on Teaching Foreign Language Performance Guidelines (1998) and Proficiency Guidelines in Speaking (1999) and Writing (2001) served as the main reference for the domain analysis for all of the foreign language exams while multiple sources of information (American Association for the Advancement of Science, 1993; Bransford, Brown, & Cocking, 1999; National Research Council, 1996, 2000, 2001, 2005) were consulted for the science domain analyses.

Finally, a college curriculum study was conducted to collect information regarding the content and skills taught in corresponding college-level courses. The college curriculum study surveyed postsecondary instructors who either nominated themselves for participation in the study or who were nominated by faculty peers because they met one or more criteria for participation. For

[2]College Board Standards for College Success have been published in English Language Arts and Mathematics and Statistics. The standards define learning objectives for six courses in middle school and high school with the intent of preparing students for AP or college-level work.

example, criteria for nomination included being an acknowledged expert in the subject area and teaching exemplary courses (see Conley, Aspengren, Gallagher, Stout, & Veach, 2006 and Conley & Ward, 2009 for the full list of nomination criteria). The instructors rated the importance of each content and/or skill in their course and also answered a variety of questions about their teaching practices, the format of their course, and uploaded documents from their course (e.g., syllabus, classroom assessments, homework assignments). These documents were reviewed by two independent raters to ensure that assertions made by instructors about the content and skills that were covered in their course were supported by the documents.

The content valued in the domain was organized and prioritized in increasing specificity starting with the big ideas of the discipline, then the enduring understandings associated with each big idea, and, finally, the supporting understandings associated with each enduring understanding (see Huff et al., 2010/this issue, for definitions of each). This approach was helpful for courses with significant amounts of content to be addressed because it provided a mechanism for organizing and prioritizing content; that is, content was only included if it related back to one or more of the big ideas in meaningful ways. Table 1 provides an example of a big idea from chemistry along with one enduring understanding and the associated supporting understandings.

To ensure that the content was not treated simply as facts to be memorized, the skills required to engage the content for deeper understanding were documented as a separate list. Table 2 provides a sample of a few of the skills that were developed in science. The skills highlighted in the table include (a) evaluating

TABLE 1
An Example Content Outline in Chemistry for one Big Idea

Big Idea: Changes in matter involve the rearrangement and/or reorganization of atoms and/or the transfer of electrons.

Enduring Understanding: Chemical changes are represented by a balanced chemical reaction that identifies the ratios with which reactants react and products form.

Supporting Understandings:

- **A.1** A chemical change may be represented by a molecular, ionic, or net ionic equation.
- **A.2** Quantitative information can be derived from stoichiometric calculations which utilize the mole ratios from the balanced equations. ***(Possible examples:*** the role of stoichiometry in real world applications is important to note so that it does not seem to be simply an exercise done only by chemists; and the concept of fuel-air ratios in combustion engines, for example, is able to provide context for this form of calculation.)
- **A.3** Solid solutions, particularly of semiconductors, provide important, non-stoichiometric compounds. These materials have useful applications in electronic technology and provide an important extension of the concept of stoichiometry beyond the whole number mole-ratio concept.

TABLE 2
Sample Skills and Skill Definitions from Science

1.	**Evaluate scientific questions**
	1A. Justification that question is in scope of investigation and domain
	1B. Evaluation and criteria for the evaluation appropriate to the question
	1C. Specification of causal mechanism(s) that is related to the question
	1D. Validity of the claim that the focus of the question is related to its purpose
2.	**Apply mathematical routines to quantities that describe natural phenomena**
	2A. Appropriateness of application of mathematical routine in new context
	2B. Appropriateness of selected mathematical routine
	2C. Correctness of mapping of variables and relationships to natural phenomena
	2D. Correctness of application of mathematical routine
	2E. Correctness of results of mathematical routine
	2F. Reasonableness of solution given the context
	2G. Description of the dynamic relationships in the natural phenomena
	2H. Prediction of the dynamic relationships in the natural phenomena
	2I. Precision of values consistent with context
3.	**Connect concepts in and across domain(s) to generalize or extrapolate in and/or across enduring understandings and/or big ideas.**
	3A. Articulation of content-specific relationships between concepts or phenomena
	3B. Prediction of how a change in one phenomenon might effect another
	3C. Comparison of salient features of phenomena that are related
	3D. Appropriateness of connection across concepts
	3E. Appropriateness of connection of a concept among contexts

scientific questions, (b) applying mathematical routines, and (c) connecting concepts in and across domain(s). In contrast, for AP courses and exams that were already primarily skill-focused, and where the pairing of content and skill was less important, the domain analysis focused on the skills to be acquired in the course. For example, students in the AP Spanish Literature course develop literary analysis skills regardless of the content of the text that is selected for reading. Thus, content outlines like those developed for science and history were not needed, although criteria were developed for selecting appropriate texts for the required list of literary readings for the course.

Capturing and maintaining the artifacts of the domain analysis (i.e., the prioritization of the content and skills) as well as the artifacts from subsequent ECD activities was a significant, nontrivial part of the process. The science SMEs used special software that facilitated conceptual (nonlinear) mapping of the big ideas, enduring understandings, and supporting understandings. On the other hand, the historians preferred to use the outline style in Microsoft Word and—still yet—the world language SMEs used Microsoft Excel to capture and refine their work. In addition, a lot of work was shouldered by the chair of each group as well as College Board staff by collecting feedback from all of the SMEs on the proposed

revisions to the artifacts, synthesizing the feedback, and presenting it back to each group for discussion, further refinement, consensus, and eventual finalization.

Domain Modeling

The domain analysis for each subject was transformed into a domain model by creating claims and evidence from the content and skills identified in the domain analysis. As mentioned, one of the primary advantages of engaging ECD is that it improves the specificity and transparency of assessment and curriculum materials that are developed. This is achieved, in part, by writing clear and concise claims about the content and skills expected of students along with documenting the observable evidence to support those claims. The claims and evidence flow from the domain analysis but are more specific in regards to what the students should know and be able to do. A key component of the process for writing claims and evidence is defining the skills in terms of the observable evidence absent specific content. In other words, skills must be defined in terms of the observable evidence or characteristics that would support inferences about student skill acquisition regardless of the specific context in which the skill is applied. These characteristics define the structure or relationships of any applicable context.

Table 2 shows a few examples of the definitions for the science skills. As shown, the SMEs provided nine different components of observable evidence for "applying mathematical routines" including, among other things, the *appropriateness of the selection and application of the mathematical routine* as well as the *correctness of the mapping of variables and relationships to natural phenomena*. These definitions are helpful because they offer a way to represent, generally, the observable characteristics of important skills, which can be used to ensure that there is consistency in the evidence associated with claims that evoke the same skill but address different content.

Two approaches were used to create the skill definitions depending on the AP subject, and these strategies impacted the directions given to SMEs to write claims and evidence. These approaches are illustrated next along with the guidelines that were provided to SMEs for writing claims and evidence. The first example is from chemistry and describes the processes that were used to write claims and evidence when the skill definitions were created before initial claim and evidence writing. Second, an example from Spanish Literature is provided to demonstrate how the process changed when the skill definitions were created after claims and evidence were initially written. A different approach was used for Spanish Literature because there was not the added component of having to integrate specific content with skills when writing claims, as in the case of science. In both cases, however, the process was iterative and required that modifications be made to the claims and evidence, as well as the skill definitions,

throughout the process. Advantages and disadvantages are discussed once the details of each are explained more fully.

Writing Claims and Evidence: An Example From Science

There were several design questions to address when writing claims and evidence including:

- What content and skill pairings are most appropriate or ideal?
- At what level of specificity (in terms of content and skill) should the claim and evidence pairs be written?
- What is the target proficiency level of the claim and evidence pair?

The science SMEs were given various guidelines to address these design questions as well as rules for addressing various semantic issues. In terms of semantics, SMEs were instructed that every claim should be clear and concise and start with the phrase, "The student can . . ." to reinforce that claims are made about what students should know or be able to do, not just about the different tasks students can perform. In addition, each claim required a verb or verb phrase that represented the skill involved in the claim (e.g., "apply"). For science, these verbs came directly from the list of skills that were considered important in the domain. It should be noted that the science skills were purposefully articulated as verbs or verb phrases so that they could be easily integrated into claims at this stage of the design. The other component of the claim was the content from the domain analysis. Because all possible pairings of content and skill were neither appropriate nor feasible given the learning goals of the AP course and the constraints of the summative exam, SMEs discussed and reached consensus regarding the most ideal pairings of content and skill. Ideal pairings were those that, for example, promoted conceptual understanding, required the student to go beyond simple rules, or promoted depth of understanding.

The grain size, or specificity, of the claim (and by extension its related evidence) was another issue that was addressed during claim writing. One general guideline given to SMEs was that the grain size of a particular claim should be such that it can be supported by a manageable amount of observable evidence. Another guideline was that if the claim provoked the question "What does that mean?" it was almost certainly too general to be meaningful. For example, the claim: "Student is able to reason scientifically" is too general and does not make clear the content with which the student is expected to interact when reasoning scientifically. On the other hand, a claim such as "Student is able to identify silicon and germanium as elemental semiconductors and be able to describe semiconductors as materials formed from the combination of two different elements" is so specific that the observable evidence that could be written is essentially a

restatement of the claim. In science, it was ultimately decided that claims should be written at the supporting understanding level, but that efforts should be taken to incorporate language that makes an explicit connection to the enduring understanding in order to keep the focus of the claim on the broader picture of the concept rather than narrowing into the specifics of the supporting information. For the same reason, there was also an attempt to include language in each claim that made a connection to or referenced in some way the big idea, especially if there were multiple parts to the big idea.

Finally, the achievement level of the claim was also addressed in the directions to the SMEs. The SMEs were told that the claims should represent summative expectations for what students should know and be able to do at the end of an AP course. More specifically, a summative expectation was defined as any claim that one would want to make about an AP student at the end of the course who deserves college credit. There are generally three levels of exam performance that are awarded credit, so claims were written to be representative of AP test scores of 3, 4, or 5 (see Plake et al., 2010/this issue, for detailed discussion of these achievement levels). These directions link directly back to the primary purpose of the AP exam as a credit-by-examination program and also assisted with resolving the issue with grain size because SMEs had to focus on writing summative claims rather than formative claims, which by their nature are more specific. The resulting number of claims that were written for each science subject was as follows: 119 for Biology, 84 for Chemistry, 107 for Physics, and 167 for Environmental Science.

Once the claims were written and endorsed by all SMEs on the panel, writing evidence began. Given that claims, by their nature, are unobservable, they need to be supported by the observable evidence that a student would have to produce to show acquisition of the claim. SMEs were instructed to start their descriptions of evidence with the phrase "The work is characterized by . . ." to reinforce the fact that evidence consists of observable characteristics of the work produced by students. The SMEs were also told that the evidence should focus on nouns (and their essential adjectives) to emphasize the notion that evidence must be concrete and observable; for example, observable evidence should not include unobservable or ambiguous phrases such as "understanding of. . . ." The SMEs were reminded not to include in the evidence any reference to the student or the task (e.g., class debate or test item), although frequently the conversation drifted in this direction. This is because SMEs were initially much more familiar with the idea of test items and classroom activities and assignments than they were with the notion of writing evidence, which required a shift in thinking and took practice. However, there is value in making this shift, as Steinberg et al. (2003, p. 30) note:

> In fact, we would argue that thinking about the relationship between what you want to observe and the way knowledge is conveyed within a domain absent the idea of a

specific type of task is a good way to broaden ideas of what can and should be considered as evidence and ways of getting it.

Finally, the SMEs were instructed to directly use the skill definitions to write evidence, so that when a particular skill was used in a claim, the evidentiary characteristics that were outlined in the definition of that skill were used as input to, and structure for, the evidence. The use of the skill definitions in writing evidence is best illustrated by an example. Consider, for instance, this sample claim from Chemistry: "Students can apply mathematics in which they evaluate the reasonableness of quantities found in stoichiometric calculations." The evidence that the SMEs wrote for this claim included:

- Correctness of chemical equation
- Correctness of chemical formulas
- Correctness of application of mathematical routine
- Correctness of coefficients interpreted as mole ratios
- Reasonableness of solution as it relates to mole ratio and differing molar masses

The skill evoked in this claim is "applying mathematical routines," the observable characteristics of which are defined in Table 2. Notice how the above evidence statements written by the SMEs parallel many of the general characteristics of "applying mathematical routines" outlined for this skill, but are more specific (i.e., "mole ratios," "molar masses") to stoichiometric calculations, which is the content referenced in this claim. In other words, the goal was to combine the content expressed in a claim with the general characteristics (evidence) listed for the skill. It should be noted that there is not a one-to-one mapping between the evidence statements for the claim and the definition for the skill because SMEs were instructed to use only those components of the skill definition that were applicable given the content of the particular claim.

The description of the process thus far implies that it was sequential; however, the actual work flow was iterative for two reasons. First, the skill definitions that were initially created were considered preliminary. As claim and evidence writing progressed, SMEs were given opportunities to modify the initial skill definitions because it was expected that changes would be needed once the SMEs began working with the skill definitions to write claims and evidence for various content topics. Up to this point, the SMEs were mainly considering the skill definitions in the absence of specific content. After modifications were made to the definitions, it was necessary to review and edit the claims and evidence accordingly to ensure consistency between the skill definitions and the claims and evidence. Second, the process of writing claims and evidence was itself iterative. There was a reciprocal relationship in that writing evidence could lead one to go

back and edit the claim, which could then lead one to go back and further edit the evidence. For example, when it was difficult to write evidence for a claim it was often because the claim needed to be revised to improve clarity or alter the grain size. Similarly, when a claim produced an enormous amount of evidence, it was often necessary to go back and divide what was probably a very general claim into more specific claims.

Writing Claims and Evidence: An Example From Spanish Literature

The process used to author claims and evidence for Spanish Literature differed from what has been described in two important ways, both of which relate to the fact that the skill definitions were drafted *after* initial claims and evidence were written. First, in terms of writing claims, this meant that there were no skills to select from to serve as the verb in the claim. Alternatively, SMEs were instructed to thoughtfully choose a verb that would best reflect the level of reasoning or skill that is expected of AP students when interacting with the content in the claim. It was emphasized that verbs such as "understands," "knows," or "comprehends" should be avoided because they are not precise enough for producing observable evidence. This is because evidence of student "understanding" or "knowing" is actually generated when a student interacts with content by identifying, describing, or evaluating a piece of content, and it those actions that should be reflected in the claim so that expectations are clear. Bloom's Taxonomy (1956) was ultimately consulted to aid verb selection, which also provided scaffolding for the SMEs to incorporate a hierarchy of skills into the claims that ranged from simple skills (e.g., identify) to more complex skills (e.g., analyze). Overall, there were 51 claims written for Spanish Literature.

The second way in which the process differed for Spanish Literature was that SMEs did not have skill definitions to shape the structure of the evidence written for each claim. Instead, SMEs were directed to think about what a student's work would need to include to demonstrate mastery of the claim, and were consistently reminded that evidence must be observable. Once the initial claims and evidence were written, the skill definitions were created by extracting and then summarizing the evidence that was articulated for the group of claims that used similar skills.

To develop the skill definitions, the SMEs had to answer the question: What constitutes similar skills? For example, if one claim used the skill *critique* and another claim used the skill *evaluate*, the SMEs needed to address whether the observable evidence for both skills was similar enough such that they could be treated as synonyms for the purposes of writing claims and evidence. This task was important because it forced SMEs to think critically about the skills they were using in their claims and to make explicit the differences between skills when needed. In the end, there were 26 skills that were used in the claims for

TABLE 3
Sample Skill and Skill Definition Categories from Spanish Literature

Category A—Identify: Recognize, refer to, name, acknowledge responses to informational questions

- recognition of key elements of a text (e.g., narrative, stylistic, linguistic, structural, rhetorical)
- recognition of relationships among concepts and among details in a text
- recognition of intertextuality (e.g., relationship between target text and another text or artistic work)
- identification of socio-cultural contexts
- identification of points of view
- identification of personal and cultural assumptions
- recognition of differences and similarities in texts, in culture, in movements
- acknowledgement of sources
- references to texts and contexts

Category B—Recount: Describe, summarize, paraphrase summary of text, paraphrase of text

- description of key elements of text (e.g., narrative, stylistic, linguistic, structural, rhetorical)
- description of contexts (e.g., historical events, dominant philosophies, social mores)

Category B

- relationship/connection between textual examples or ideas/concepts/themes and written, audio, visual, or audiovisual material
- formulation of oral and written questions about texts and contexts to demonstrate understanding
- main ideas/themes, supporting points in written, audio, visual, or audiovisual material
- description of relevant information from other disciplines
- presentation of information in a descriptive form

Spanish Literature, which were organized by synonym into five categories representing a range of skills from simple to complex. See Table 3 for a sample of the first two skill categories.

Before finalizing the skill definitions, the SMEs needed to determine whether the general evidence that was extracted from individual evidence statements included everything that was important for the skill as well as whether anything extracted needed to be revised or clarified. In addition, any inconsistencies and redundancies needed to be resolved as well. Iterations between the skill definitions and the claims and evidence were needed because any changes to the skill definitions needed to be reflected in the claims and evidence and vice versa.

Although both approaches required that the work be iterative, developing skill definitions based on preliminary claims and evidence, as for Spanish Literature, was generally an easier approach than the one used in the science subjects because of the difficulty involved for SMEs to think about observable evidence for skills absent of content. Nonetheless, the advantage of requiring SMEs to define skills upfront, as in the case of science, was that it provided helpful and specific guidelines for writing claims and evidence. These guidelines were particularly useful for the domains that sought to focus equally on content and skills.

DISCUSSION

The approach that was used to determine the content and skills to be covered in AP courses and measured on the exams parallels traditional methods in several ways. Most notably, it relied on the judgment of SMEs, as well as the review and use of applicable curriculum frameworks, national standards, and learning science research. Curriculum studies were also conducted. However, the task of articulating the required content and skills only began with this work as opposed to ending with it. The processes were extended by using backwards design to organize and prioritize the content and skills in the domain, and then using ECD to specify the claims and evidence to represent the summative learning goals for the course and exam.

Challenges to Using ECD

There are several challenges to articulating the domain in terms of claims and evidence. First, lack of time to complete work during face-to-face meetings was frequently an issue. The start-up time required to orient SMEs to ECD and the iterative nature of the work added strain to the project timeline and made the work resource-intensive. In some cases, SMEs viewed iterations as having to do things over, thinking it must have been initially done the wrong way. To mitigate these concerns, the iterative nature of the work must be emphasized from the beginning. Because work was frequently not completed during face-to-face meetings, SMEs sometimes wrote claims and evidence individually as homework, despite the general preference to work as a group and with the help of a facilitator. When SMEs worked individually, additional steps were needed to have the work synthesized, presented to the group for discussion, revision, and eventual endorsement.

A related challenge was that the work itself was demanding and required a steep learning curve. It was difficult for SMEs to think in terms of observable evidence as opposed to tasks, which is what they were more comfortable thinking about and where the discussions frequently focused. Another challenge was defining the appropriate level of specificity at which to write the claims and evidence. If the grain size was too small, the result was claims that functioned more as formative claims rather than as the desired summative claims. If the grain size was too large, the resulting claims were so broad that their usefulness in informing both curriculum and assessment design was questionable. An area for future research would be to address ways to better define the appropriate grain size in a way that is transparent for the SME who is familiar with the content and skills, but new to writing claims and evidence. Even once defined, however, experience suggests that it will remain a difficult task to consistently write claims at a specified grain size, so time should be built into the process for adequate review, discussion, and revision.

Finally, this work involved forging new territory. The very act of articulating claims and evidence for these disciplines actually helped further understanding within the disciplines themselves. As a field, our understanding about how deep conceptual understanding and complex reasoning skills are acquired and evidenced in specific subject areas is still evolving. As Huff et al. (2010/this issue) indicate, the SMEs involved in the domain modeling had to rely largely on their own expertise when selecting the optimal combinations of content and skill or defining observable evidence for the claims because this information simply did not exist de facto.

Advantages of Using ECD

Despite the challenges, there are several advantages to articulating the domain in terms of claims and evidence. A key advantage is that it provides a foundation for making stronger links between curriculum, instruction, and assessment, which is essential for the AP Program. Because the claims are clear statements of the knowledge and skills that students should acquire and are accompanied by specific evidence to support each claim, the expectations for what should be taught in the course and assessed on the exam are transparent. There is no guesswork involved on the part of the teacher or the item writer about what is valued in the domain or what characteristics of student work are required for evidence. In addition, defining the skills in terms of observable evidence ensures that the skills are not ignored and that teachers and test developers become forced to look for evidence of not only content, but for skill as well.

As others have noted (Steinberg et al., 2003), evidence of the skill integrated with the content is a critical piece that is missing from traditional learning goals. Having evidence that describes this integration provides teachers with guidance for designing instruction that develops both the content and skill in the context of each other. The ECD artifacts provide a level of detail that supplies teachers with clear targets for instruction as well as assessment because the activities involved in each emphasize the importance of defining the targeted understanding in a way that captures the use of the knowledge, and not just discrete concepts or facts. Clear learning goals in the form of claims and evidence can be beneficial to students as well because students will know exactly what is expected of them. This can foster metacognition, if instruction permits, because the students will know their goals and can reflect on their understanding of the goals (Bransford et al., 1999).

CONCLUSIONS

The validity argument for test score interpretation is strengthened through ECD. Kane (2006) suggests that the act of validating an assessment includes two types

of arguments: an interpretive argument and a validation argument. The interpretive argument lays out the proposed test score interpretations and uses while the validity argument provides an evaluation (often empirical) of what is proposed. In Kane's framework, the interpretive argument begins with linking test performance to test scores (p. 27): "Initial inference in a quantitative interpretive argument is to be from a record of performance on some task (datum) to a score (the claim)." By using ECD to articulate the domain, the link between performance on tasks and test scores is robust because there is a well-documented, explicit rationale for why specific tasks were designed for the test in the first place: the claims and evidence and their connections to tasks and test specifications (Hendrickson et al., 2010/this issue).

REFERENCES

American Association for the Advancement of Science. (1993). *Benchmarks science literacy*. Washington, DC: Author

American Council on Teaching Foreign Language. (1998). *ACTFL performance guidelines for K–12 learners*. Alexandria, VA: Author.

American Council on Teaching Foreign Language. (1999). *ACTFL proficiency guidelines— Speaking*. Alexandria, VA: Author.

American Council on Teaching Foreign Language. (2001). *ACTFL proficiency guidelines— Writing*. Alexandria, VA: Author.

Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook 1: The cognitive domain*. New York: David McKay Co Inc.

Bransford, J., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Conley, D., & Ward, T. (2009). *College curriculum study in world languages and literatures: Final report*. Eugene, OR: Educational Policy Improvement Center.

Conley, D., Aspengren, K., Gallagher, K., Stout, O., & Veach, D. (2006). C*ollege Board Advanced Placement® best practices course study*. Eugene, OR: Center for Educational Policy Research.

Hendrickson, A., Huff, K., & Luecht, R. M. (2010/this issue). Claims, evidence, and achievement-level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education*.

Huff, K., Steinberg, L., & Matts, T. (2010/this issue). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education.*

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement*, *4th edition* (pp. 17–64). Washington, DC: American Council on Education.

Mislevy, R. J., & Risonscente, M. M. (2005). *Evidence-centered design: Lays, structures, and terminology* (PADI Technical Report 9). Menlo Park, CA: SRI International and University of Maryland.

Mislevy, R. J., Almond, R. G., & Lukas J. (2004). *A brief introduction to evidence-centered design.* (CRESST Technical Report 632). Los Angeles: Center for the Study of Evaluation, CRESST, UCLA.

National Research Council. (1996). *National Science Education Standards*. Washington, DC: National Academy Press.

National Research Council. (2000). *Inquiry and the National Science Education Standards*. Washington, DC: National Academy Press.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2005). *Systems for state science assessment*. M. W. Bertenthal & M. R. Wilson (Eds.). Washington, DC: Committee on Test Design for K–12 Science Achievement.

Plake, B. S., Huff, K., & Reshetar, R. (2010/this issue). Evidence-centered assessment design as a foundation for achievement-level descriptor development and for standard setting. *Applied Measurement in Education.*

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement*, *4th edition* (pp. 307–353). Washington, DC: American Council on Education.

Steinberg, L. S., Mislevy, R. J., Almond, R. G., Baird, A. B., Cahallan, C., DiBello, L. V., et al. (2003). *Introduction to the Biomass project: An illustration of evidence-centered assessment design and delivery capability* (CRESST Technical Report 609). Los Angeles: Center for the Study of Evaluation, CRESST, UCLA.