Evidence-Centered Design: Recommendations for Implementation and Practice

Amy Hendrickson, Maureen Ewing, Pamela Kaliski
The College Board

Kristen Huff
Regents Research Fund

Abstract

Evidence-centered design (ECD) is an orientation towards assessment development. It differs from conventional practice in several ways and consists of multiple activities. Each of these activities results in a set of useful documentation: domain analysis, domain modeling, construction of the assessment framework, and assessment implementation/delivery (Mislevy & Haertel, 2006).

In this article we focus on the four primary challenges we have encountered in our work with ECD in the context of large-scale educational assessment. For each challenge, we identify potential mitigation strategies as well as research studies or other endeavors that we think are helpful in advancing the science of ECD. The challenges discussed are: integrating learning theory into assessment design; identifying the appropriate levels of specificity with which to document the claims and evidence; developing and evaluating task models; and strategically incorporating iteration into the design process.

Keywords: Evidence-centered design, Large-scale assessment, Assessment design

Evidence-centered design (ECD) is an orientation towards assessment development. It differs from conventional practice in several ways: (a) the amount of work required up front in the design phase (i.e., before items are written); (b) the prioritized role of observable evidence in design and development; and (c) the documentation and use of *claims*, *evidence*, and *task models* (Huff, Steinberg, & Matts, 2010; Mislevy, Steinberg, & Almond, 2003). Claims are an articulation of measurement targets that reflect what is valued in the domain and are shaped by the purpose of the assessment. The evidence required to support claims are our assumptions and hypotheses about how these latent proficiencies (i.e., the measurement targets) manifest in student work. Task models are a collection of task features that define the task according to the evidentiary focus and intended cognitive demand.

Evidence-centered design consists of multiple activities. Each of these activities results in a set of useful documentation: domain analysis, domain modeling, construction of the assessment framework, and assessment implementation/delivery (Mislevy & Haertel, 2006). It is important to note that ECD is iterative and incremental; as a result, it is somewhat arbitrary when the artifacts are drafted and finalized. (See Figure 1 for a depiction of the iterative process of ECD.)

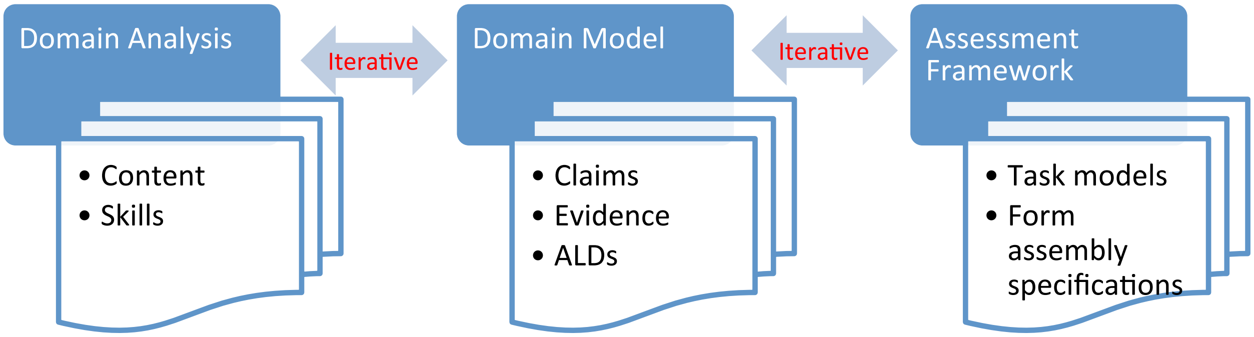

Figure 1.

The Iterative Evidence Centered Design process (Adapted from Huff, Steinberg, & Matts, 2010)

For example, with the College Board's Advanced Placement (AP®) program, each activity and the resulting artifacts were as follows:

• Analyzing the domain, which resulted in prioritization of content and skills, organized conceptually to facilitate curriculum and assessment design to support deep understanding.

• Modeling the domain, which resulted in claims and corresponding evidence. Claims and evidence emerge directly from the content and skills produced by the domain analysis. (Claims are also referred to as Learning Objectives in AP publications. For example, see College Board 2011, 2012.)

• Articulating the achievement level descriptors that identify what examinees at each of the achievement levels are expected to know and be able to do.

• Constructing the assessment framework that resulted in task models, in some content areas, and form assembly specifications.

In brief, ECD is a set of activities and artifacts that facilitate explicit thinking about (a) what content and skills are both useful and interesting to claim about examinees given the purpose of the assessment; (b) what is the reasonable and observable evidence in student work or performance required to support the claims; and (c) how tasks (items) can be developed within the constraints of the assessment.

Quality assessment design and development is challenging work. What is the added benefit of using ECD? Not only is it more work; it is more challenging work. Although true on both counts, the additional costs are far outweighed by the major benefits that (1) assessments better reflect and measure what is taught and valued in the classroom, and (2) resulting score inferences are strongly supported by an evidentiary argument. Given the increasingly high stakes attached to educational achievement tests (especially teacher effectiveness evaluations), these two benefits alone reflect the emphasis of ECD on what is taught and valued in the classroom. Moreover, these benefits justify the increased resources required at the beginning of the design endeavor. Other important test development benefits include improved construct equivalence across forms within and across years, as well as the potential to create many more items in a given period of time once robust task models are available. Benefits to teachers include a clear and detailed articulation of what students should know and be able to do; explicit alignment of curriculum, exam design, item coding, rubric design, and score reporting; use of item types that consistently reflect the behaviors of practitioners in the discipline; a process for articulating, validating, and reporting what students at each achievement level in the discipline know and are able to do. In whole, ECD is practical, feasible, and replaces the traditional test development method that relies heavily on post-administration evaluation (i.e., item analysis, equating and generation of achievement level descriptions).

Currently, the use of ECD is quite limited (e.g., Cisco [Behrens, Mislevy, Bauer, Williamson, & Levy, 2004]; Advanced Placement [see *Applied Measurement in Education,* 2010 *23*(4)]; Common Core assessment consortia [see The Partnership for Assessment of Readiness for College and Careers (PARCC), 2012]). We hypothesize, however, that the use of ECD will become less resource intensive once it is employed more broadly in assessment design and development, and we collectively identify pathways through the current challenges. We address four of these challenges in this article.

To communicate the common challenges and their possible mitigation strategies, in this article we focus on the four primary challenges we have encountered in our work with ECD in the context of large-scale educational assessment. For each challenge, we identify potential mitigation strategies as well as research studies or other endeavors that we think are helpful in advancing the science of ECD. The challenges discussed are: integrating learning theory into assessment design; identifying the appropriate levels of specificity with which to document the claims and evidence; developing and evaluating task models; and strategically incorporating iteration into the design process.

Challenge #1: Integrating Learning Theory into Assessment Design

It is essential to articulate our theory about how students learn and develop mastery in a given domain early in the assessment design process. This theory defines our approach to cognitive complexity and achievement level descriptions (ALDs), and it is the key to ensuring that the assessment measures what is valued in teaching and learning (Huff & Plake, 2010). Cognitive complexity addresses (a) the skills, processes, and/or practices that we value in this content area; (b) how these factors change as students become more proficient; and (c) the nature of the interaction with content vis-a-vis cognitive demand. Our approach to cognitive complexity directly affects the claims we make, the evidence to support the claims, our achievement level descriptions, and ultimately, the tasks we develop as well as the characteristics of the score scale (Hendrickson, Huff, & Luecht, 2010). It has been more than a decade since the National Research Council (NRC) published seminal works on the role of learning theory in contemporary assessment design (NRC 2000, 2001). However, the practice of incorporating these theories is nascent at best.

The first step in meeting this challenge is to realize and acknowledge that there is a role for theories about how students learn and develop mastery within a given domain in assessment design. If the role is understood and valued, then explicit steps should be integrated into the design process from the beginning to articulate and refine our theories of knowing and learning in the domain. The most effective mitigation strategy is ensuring that the right talent is a part of the assessment design team from the beginning—namely those who understand how students learn and develop deep conceptual knowledge in the target domain.

Much more research on how students learn and develop mastery in a given domain is needed. Such research is most effective when designed and executed by a partnership of learning scientists, cognitive scientists, and assessment designers. Similarly, more "think-aloud" and cognitive labs should be conducted with assessment tasks throughout the design process to provide opportunities to refine our hypotheses about the knowledge and skills students are employing to engage successfully with tasks. However, cognitive labs are typically resource-intensive and costly, which is a likely reason that they are not conducted more often in current practice. As such, more streamlined approaches to incorporating cognitive labs into assessment design are needed, such as a more "focus-group" approach rather than lengthy analysis of qualitative verbal report data collected through cognitive labs. For example, rather than conducting many cognitive labs, transcribing all of the verbal reports, developing a detailed coding scheme, and having at least two coders code all verbal reports, the process could be simplified by having subject matter experts (SMEs) simply read the verbal reports and decide whether or not the appropriate skill is being elicited.

Challenge #2: Identifying Appropriate Levels of Specificity in the Claims

Writing claims and evidence at the appropriate level of specificity, or grain size, is another challenge to consider. Generally speaking, the grain size of claims and evidence should be written at a level that aligns to the purpose of the test. Even within the boundaries of a particular test purpose, however, questions about *appropriate* grain size may still arise. General rules of thumb to address grain size issues have been proposed (Ewing, Packman, Hamen, & Thurber, 2010). However, these guidelines address only the grain size of claims and evidence regarding a single phase of ECD (i.e., the domain model)— not in terms of how the grain size of claims and evidence impacts subsequent phases of ECD (i.e., the Achievement Level Descriptors [ALDs], the assessment framework, and task model development). The challenge, therefore, is twofold: (a) providing information to the subject matter experts (SMEs) early about the future use of the claims and evidence in creating the ALDs and task models and (b) being able to give the SMEs useful guidelines regarding the level at which to write claims and evidence such that they are useful for students and teachers to focus on what is important in the domain and understand the nature of the performance the student should aim to achieve, as well as to yield robust task models.

In the simplest sense, a grain size issue emerges when a claim is either too broad or too narrow. For example, the claim that a "Student is able to reason scientifically" is broad and makes generating appropriate evidence difficult. Conversely, a claim that a "Student is able to identify silicon and germanium as elemental semiconductors" is so specific that any observable evidence that could be written is essentially a restatement of the claim. A general guideline, therefore, is to think about the grain size of a particular claim in light of the generated evidence. In other words, a claim should be written to a grain size such that it can be supported by a "manageable amount" of observable evidence. Evidence that is too lengthy or unfocused suggests that the claim is too broad and that it should be broken into one or more claims with the evidence revised accordingly. Evidence that is simply a restatement of the claim, suggests that the claim is too narrow.

The process of evaluating whether a claim is at the right grain size by determining whether it is associated with a manageable amount of evidence involves a degree of subjectivity. If instead we were able to frame one or more guidelines in terms of what would be beneficial for task model development, we may be able to create more objective criteria for writing claims and evidence at the appropriate grain size. For example, it is unlikely that a robust task model can be developed with evidence that is essentially a restatement of the claim (i.e., multiple items cannot be generated). The question, then, is whether other features of claims and evidence that affect the robustness of task models can be identified and documented for purposes of improving claim and evidence writing. Regarding the challenges within task model development, one feature relates to the claim and its associated evidence. For example, consider a claim that is written at the highest level of achievement (e.g., the advanced level). There are two scenarios possible for the evidence: (a) all of the associated evidence is written/judged to be at the same level of achievement (i.e., advanced) or (b) components of the evidence span more than one level of achievement (i.e., basic, proficient, and advanced). Either approach is technically feasible, and both approaches have been taken within various disciplines for AP. Some disciplines may, in fact, have strong preferences for one approach over another. However, it is important to acknowledge early in the design work how the two approaches may affect task model development in terms of both the number and robustness of task models. This issue is discussed further under Challenge #3.

One mitigation strategy that holds promise for generating claims and evidence at a grain size useful for task model development is to have the same SMEs who write the claims and evidence develop at least a few task models for a small sample of claims and evidence. This should occur early in the process before all of the claims and evidence are drafted so that lessons learned can be incorporated into the claim and evidence writing efforts. Giving SMEs the opportunity to draft task models will help them gain a fuller appreciation for the entire process, begin to get a feel for how claims and evidence feed into task models, and possibly help them to identify features of claims and evidence that yield useful task models.

Research studies should be conducted on how to best support SMEs to consider the interplay between claims, evidence, and task models early in the process and whether this contributes to the overall process results. Research studies should also be conducted to help identify the characteristics of claims and evidence that are likely to yield useful task models.

Challenge #3: Developing and Evaluating Task Models in Evidence-Centered Design

The importance and process of task model[1] development has been described and discussed by many assessment design researchers and practitioners (e.g., Hendrickson, Huff, & Luecht, 2010; Huff, Alves, Pellegrino, & Kaliski, in press; Mislevy & Haertel, 2006). Yet, successful task model development during the establishment of an assessment framework remains a significant challenge. First, there are few guidelines for informing task model development and the desired characteristics, such as: (a) Is the task model at the right grain size or sufficiently detailed? (b) When is the task model complete? (c) Is the task model robust (i.e., does the task model allow for generation of an appropriate number of items)? Second, there are no criteria in place to evaluate the completed task models and determine whether the items generated by a given task model are functioning well.

We suggest two mitigation strategies to assist with this two-fold challenge of developing and evaluating task models. The first is a straightforward approach to task model development in which SMEs are instructed to complete a series of steps for a given claim. The SMEs'

[1] Note that in some circles, the terms "task model" and "item template" are used interchangeably, whereas in other circles task models are more general than item templates. This inconsistent language has caused some confusion in the task model literature, but this is not the focus of the current challenge; therefore, we will discuss only task model development in this section.

responses to these steps are molded into task models by the assessment designers. The second strategy is a task model evaluation criteria framework proposed to help test developers determine whether task models are functioning well. This strategy is an elaboration of the mitigation strategy mentioned in the previous section regarding the grain size of claims and evidence statements because this mitigation strategy serves the purpose of both informing the grain size of claims and evidence and assisting with task model development.

Straightforward Task Model Development

In this approach, SMEs are *not* instructed specifically to create task models. Rather, they are instructed to complete a series of steps for a given claim that is a target of measurement on the assessment being designed. The SMEs' responses to these steps are molded into task models by the assessment designers.

The first step is to *remove all ambiguity in the claims and redefine this as observable evidence.* Ideally this step has already been done while establishing the domain model (Ewing, Packman, Hamen, & Thurber, 2010). If, however, this step has not been done (i.e., the claims still include terms such as "students understand…" or the student can produce a work product that is "comprehensive"), then more work is needed. Subject matter experts will need to ask themselves what it means for a student to understand something (i.e., how "understanding" is observed) and what it means for a work product to be comprehensive (i.e., what we need to see to know that the work product is comprehensive). These are the types of questions in which the assessment design team engages throughout the ECD process. The answers to these questions are used to articulate the evidence statements and, when necessary, revise the claims (Ewing et al., 2010). The benefit of this step for task model development is that the answers to these questions are fodder for the manipulable task features (i.e., the variables and the values) that will be articulated in the task model. For example, in Figure 2, number 3 under "Manipulable features of complexity," "Type of statement/alternative" is listed as a manipulable feature that could change within this task model and may have originated from ambiguity in the original claims (e.g., what specifically the student is being asked to do).

Construct Identifier:	Biology
Primary Context:	Cell division
Competency Claim	The student can construct explanations for how DNA is transmitted to the next generation via the processes of mitosis, meiosis, and fertilization.
Proficiency Level	Basic
	Evidence Documentation
1.	Description of the purpose of mitosis and meiosis
2.	Description of the products of mitosis and meiosis
3.	Description of the behaviour of the chromosomes during the phases of mitosis and meiosis
4.	Explanation of the processes of mitosis and meiosis
5.	Comparison and contrast of the processes of mitosis and meiosis
6.	Use and recognition of vocabulary specific to cell division
	Manipulable features of complexity
1.	Type of cell division (mitosis is simpler than meiosis)
2.	Number of steps in process (mitosis has fewer steps than meiosis)
3.	Type of statement/alternative (definition is less challenging than explanation)
4.	Use of vocabulary particular to cell division will increase complexity: ploidy, tetrads, synapsis, crossing over, sister chromatids, homologous chromosomes, segregation, equatorial plate, cytokinesis
5.	Phase of cell division in question; the events in some phases are more conceptually difficult than the events of other phases
6.	Making a comparison (more challenging) vs. selecting a true statement (less challenging)
	Features irrelevant to complexity
1.	Number of chromosomes in a cell
2.	Type of organism in which the processes occur

Figure 2.[2]

Example task model for a biology cell division claim (Adapted from Huff, Alves, Pellegrino, & Kaliski, in press)

[2] This work was sponsored as a research project, and no items produced by this project were or will be included on actual AP examinations. Example task models shown in this manuscript do not represent actual AP task models or task model templates, and are simply used for illustrative purposes to explain components of a task model. This work was conducted simultaneously as the completion of the final NSF report, and as such the reader may encounter edited versions of the claims and evidence in the final NSF report. The AP Biology claims and evidence in both the final NSF report and this manuscript represent earlier versions of the materials used for current AP Biology course and assessment materials.

The second step is to *state intended cognitive complexity* of the claim to determine where, within the levels of achievement, the task models assessing this claim will be situated. Once the level of achievement of a claim is decided, the level of achievement of the task model is also determined given that the task model emerges from the evidence statements for a given claim. Ordered task models are the result. In Figure 2, the Proficiency Level cell indicates that the claim is targeting the "Basic" level; thus, any resulting task models should also target the Basic level.

As mentioned in the previous challenge on grain size, however, it is possible to have evidence statements within a claim that are targeted at more than one achievement level. The grain size example considered a claim intended to assess the highest achievement level, but with some evidence statements written to lower achievement levels. In this case, it is acceptable to indicate the intended level of cognitive complexity at the evidence statement level, rather than the claim level. In Figure 2, this would mean that the Proficiency Level would be indicated after each evidence statement, rather than associated with the overall claim. Separate task models would then be created for evidence statements that are targeted to different levels of achievement to maintain the idea of ordered task models.

One benefit of ordered task models is the reconciliation between the unobservable construct being measured and the psychometric properties of the scores from the assessment. If, when claims and evidence are written, they are targeted at a specific achievement level, then task models developed for a certain claim or evidence statement will be nested within an achievement level. In this scenario, score interpretation for the assessment is richer and is supported by a stronger validity argument.

The third step in task model development is to *articulate the features that impact the complexity of the tasks that elicit the evidence for a given claim*. In other words, identify sources of cognitive complexity. For example, consider the Biology claim in Figure 2: "The student can construct explanations for how DNA is transmitted to the next generation via the processes of mitosis, meiosis, and fertilization." This claim is targeted at a "Basic" level of achievement. One source of cognitive complexity is "Type of cell division"; specifically, tasks that ask about meiosis are more complex than tasks that ask about mitosis. Identifying sources of cognitive complexity is not a typical step in test development (Schmeiser & Welch, 2006); however, much research has been conducted in this area (e.g., Kaliski, Huff, & Barry, 2011; Schneider, Huff, Egan, Tully, & Ferrara, 2010). When sources of cognitive complexity are identified, test developers can use them to create items that cover the range of ability within a given achievement level or to create separate task models for each targeted level of complexity. For example, because tasks that ask about meiosis are more complex than tasks that ask about mitosis, the test developers should consider if two separate task models should be created—one for mitosis and one for meiosis.

We hypothesize that going through these steps better aligns with the typical work of the subject matter experts as compared to asking them to identify variables and values related to work products, stimuli, and constraints on complexity and evidentiary focus. (See Hendrickson, Huff, & Luecht [2010] for more information on these topics.)

Evaluation Criteria Framework for Task Models

Another proposed mitigation strategy for this twofold challenge of developing and evaluating task models is to develop an evaluation framework for task models. An evaluation framework would help assessment designers determine whether their task model development efforts are successful. See Table 1 for the proposed evaluation framework. We will refer to the example task model template shown in Figure 3 to help articulate the pieces of this evaluation framework.

Table 1.

Framework for task model evaluation criteria

1. **Criteria related to construct-relevance, specificity, and scalability**
   a. Is there more than one evidentiary foci for the task model?
      - If yes, then consider separate task model for each evidentiary foci.
      - If no, create one task model.

   b. Are the evidence statements targeted to more than one level of cognitive complexity?
      - If yes, then consider separate task models for each level.
      - If no, then keep the task model(s) at one level of complexity.

   c. Is there evidence of scalability for the task model?
      - If yes, then the task model is at an acceptable grain size.
      - If no, then the task model is at too specific of a grain size.
2. **Criterion related to item statistics and intended cognitive complexity**
   a. Is there a large range of variability in the item difficulty statistics for items written to a particular task model?
      - If yes, is a small proportion of items causing the large range?
        - If yes, then the task model is acceptable, but the items may need to be revised.
        - If no, consider revising the task model so that the majority of the items are better aligned with the intended cognitive complexity level.
      - If no, then do the item difficulty statistics align with what is expected given the intended cognitive complexity level?
        - If yes, then the task model is acceptable.
        - If no, consider revising the task model so that the majority of the items are better aligned with the intended cognitive complexity level.

The evaluation criteria framework has two components: (a) criteria related to construct relevance, specificity, and scalability, and (b) criteria related to item statistics and intended complexity. Under the first component, after a task model is drafted, three questions must be asked. Question 1 is “Is there more than one evidentiary foci for the task model?” In other words, is there more than one evidence statement? If so, do these evidence statements call for different types of tasks? If the answer is yes, then the SME or test developer should consider separate task models for each evidence statement. If the answer is no, then one task model should suffice to elicit the intended evidence. Question 2 is “Are the evidence statements targeted to more than one level of cognitive complexity?” If the answer is yes, then the SME or test developer should consider a separate task model for each level. If the answer is no, then keep the task model at one level of complexity. Finally, Question 3 asks, “Is there evidence of scalability for the task model?” Scalability is related to the number of variables with manipulable features in a task and the number of values that can be plugged into these features, as well as the variety of evidentiary foci. Thus, a task model that has high scalability will be robust and can generate many items through the possible values. For example, Figure 3 depicts a task model with high scalability. Although there are only two options for the item stem, the various values in the item distracters and response options allow for many item possibilities.

Primary Context: Cell division

Competency Claim: The student can construct explanations for how DNA is transmitted to the next generation via the processes of mitosis, meiosis, and fertilization.

Key: A	***Intended Level of Complexity***: Moderate

Stem: Which of the following statements best illustrates one aspect of the **[type]** of meiosis?

A. [Key]
B. [Distractor 1]
C. [Distractor 2]
D. **[Distractor 3]**

Stem elements that are allowed to vary:
[type] range: "definition", "description"

Allowable ranges for key and distractors:

[Key]
1. By the end of telophase I, each pole of the cell has a haploid set of chromosomes.
2. By the end of anaphase II, each pole of the cell has one half of a set of chromosomes.
3. During anaphase II, the centromere splits and each sister chromatid moves to the opposite pole of the cell.
4. DNA is replicated once before meiosis and divided twice during meiosis.
5. Telophase II is followed by cytokinesis.
6. During prophase I, tetrads form.
7. During anaphase II, sister chromatids segregate.
8. During anaphase I, homologous chromosomes segregate.
9. During prophase I, crossing over occurs and alleles are exchanged.
10. During metaphase I, chromosomes line up in homologous pairs at the centre of the cell.
11. During anaphase I, the two chromosomes that form a tetrad move to opposite poles of the cell.
12. During prophase I, each individual chromosome has two sister chromatids joined by a centromere.

[Distractor1]
1. Meiosis alone ensures that genes are passed from one generation to the next generation.
2. During metaphase II, sister chromatids segregate.
3. During metaphase II, homologous chromosomes segregate.
4. During metaphase I, crossing over occurs and alleles are exchanged.
5. During metaphase I, chromosomes form a single line at the centre of the cell.
6. During metaphase II, chromosomes line up in pairs along the centre of the cell.
7. During metaphase II, crossing over occurs and some genes are exchanged between homologous chromosomes.

[Distractor2]
1. Fertilization alone ensures that genes are passed from one generation to the next generation.
2. After telophase I, DNA duplicates to prepare the cell for the second meiotic division.
3. During prophase II, chromosomes form homologous pairs.
4. During prophase II, tetrads are formed and crossing over occurs.
5. During prophase II, crossing over occurs and alleles are exchanged.
6. During metaphase II, crossing over occurs and alleles are exchanged.
7. During prophase II, chromosomes line up in pairs along the centre of the cell.
8. During interphase I, DNA duplicates to prepare the cell for the second meiotic division.

[Distractor3]
1. During prophase I, each individual chromosome has four sister chromatids joined by a centromere.
2. Telophase II follows cytokinesis.
3. DNA is replicated once before meiosis and divided once during meiosis.
4. DNA is replicated twice before meiosis and divided once during meiosis.
5. During anaphase II, chromosomes line up at the equatorial plate of the cell.
6. During telophase II, DNA replicates to prepare for the next meiotic division.

Constraints:
[TYPE]=1 & [Key]<4 ||[TYPE]=2 & [Key]>3
[Key]<4 & [Distractor1]<2 || [Key]>3 & [Distractor1]>1
[Key]<4 & [Distractor1]<2 || [Key]>3 & [Distractor1]>1
[Key]<4 & [Distractor2]<2 || [Key]>3 & [Distractor2]>1
[Key]<4 & [Distractor3]<2 || [Key]>3 & [Distractor3]>1
[Key]>5 & [Distractor3]>4 || [Key]<6 & [Distractor3]<5
[Key]<6 & [Distractor2]<3 || [Distractor2]>2||[key]>5
[Key]=5 & [Distractor3]=2|| [Distractor3]!=2

Figure 3.[2]

Example Task Model Template for Generating Moderately Challenging Items on Meiosis (Adapted from Huff, Alves, Pellegrino, & Kaliski, in press)

The second component of the evaluation framework is related to item statistics and intended complexity. This component becomes relevant once pilot performance data are available for items that were generated from a task model. When items are intended to assess a particular level of achievement, the item difficulty statistics should align with the intended level. For example, if an item is written to assess the highest level of achievement, the item's statistics should generally indicate that this is among the more difficult items. While this will never be a perfect relationship, for a variety of reasons, the item difficulty statistics should generally be in alignment with the intended level of cognitive complexity. The primary question a test developer must ask is as follows: "Is there a large range of variability in the item difficulty statistics for items written to a particular task model?" If the answer is yes, the next question the test developer should ask is, "Is a small proportion of items causing the large range?" If a small proportion of items is causing the large range, then the task model is acceptable as is. However, these outlying items may need to be revised.

If, on the contrary, a small proportion of items cannot be identified, then the test developer should consider revising the task model so that the majority of the items developed from the task model are better aligned with the intended cognitive complexity level. If the answer to the primary question of whether or not there is a large range of variability in item difficulty estimates is no, the next question the test developer should ask is as follows: "Do the item difficulty statistics align with what is expected given the intended cognitive complexity level?" If they do, then the task model is acceptable as is; however, if they do not, then the test developer should consider revising the task model so that the majority of the items are better aligned with the intended cognitive complexity level. The identified manipulable features of complexity should be leveraged when revising a task model to better align it to a level of intended cognitive complexity.

As proposed earlier, research studies should be conducted on how to best support subject matter experts in understanding the interplay among claims, evidence, and task models early in the assessment development process and whether this benefits the overall process results. Research studies should also be conducted to help identify the characteristics of claims and evidence that are likely to yield useful task models. Further research should be conducted to develop and test the proposed mitigation strategies for implementing task model development. Subject matter experts should be convened for a pilot task model development workshop and should provide feedback on the process, including the ease of the procedures and whether the approach is feasible.

Research should be conducted to test the utility of the evaluation framework. This research will require that pilot data be collected from a series of items generated from some task models. Test developers should answer the questions in the proposed evaluation framework and determine whether these questions lead to useful conclusions about the quality of the task models. Another potentially useful research study would be to administer sets of items that represent different instantiations of one task model, in which only one hypothesized manipulable feature of cognitive complexity is varied. Then, the item statistics from these item instantiations would be compared and would help to inform possible revisions to the task models in cases where the item statistics and intended levels of cognitive complexity do not align.

Challenge #4: Creating Points of Iteration within the Evidence-Centered Design Process

Although each stage of the ECD process is encountered in turn, it should not be considered a linear process; rather, iteration must occur within and across each step of the process. (See Figure 1.) As ECD becomes integrated into a testing program, much of the time for iteration can be planned ahead and built into the process. However, the need for additional unplanned time and resources to revisit previous discussions and decisions is a greater challenge while implementing ECD. These iteration steps extend the required time of the process, incur additional cost, and may give the appearance that the "The developers/implementers implementers didn't know what they were doing" or "They did it wrong the first time." Iteration IS an important part of the process that should be emphasized from the start of the process. Several process points when iteration is most likely needed and productive are described below.

As the brainstorming and development work within each step of the ECD process occurs (e.g., domain analysis, domain model), it is natural that some discussions lead to a "circling back" to previous discussions and decisions. For example, as it is decided that one topic is a Big Idea and another an Enduring Understanding, the other Big Ideas and Enduring Understandings should be reviewed to ensure that the Big Ideas really are big and that there is a balance across the Big Ideas. (See Ewing, Packman, Hamen, & Thurber [2010] for definitions and details on these concepts.) This review allows a revisit of the level of specificity (i.e., grain size) issue discussed earlier.

Similarly, as one moves through the steps of the ECD process (e.g., from the domain analysis to the domain model) and more decisions are made, it should be expected that previous decisions need to be "verified". As described for the previous two challenges, the claims and evidence should be written with mindfulness of their future use as the basis for ALDs and task models. Even with this forethought, it is likely that as one moves from the domain analysis to the domain model, necessary modifications to Big Ideas and Enduring and Supporting Understandings will be revealed. Similarly, as the test specifications are developed, holes in the domain model may be uncovered that require subject matter experts to reconsider decisions made during the domain analysis or modeling stages. For example, through the process of formalizing test specifications, domain experts decide on desired weights for the claims on the assessment. This process may result in identification of changes needed to the domain model.

Another possible time for iteration is for stakeholder validation of any and/or all of the products of the ECD process. For example, it may make sense to have an external validation, approval, or sign-off of the results of the domain analysis as these elements form the basis for all other products of the process. If this external validation reveals suggested changes, time and resources must be built in to make those changes. Making changes at this stage of development, as soon as the domain analysis is completed, may save the time/effort of re-work further along the development path.

The AP Program has had an extended timeline for the development of its first ECD redesigned courses and exams. When the AP course and exam were revised, the implementation of ECD for a large-scale, high-stakes exam had not been tried previously. Therefore, timelines were based on previous experience and timeframes of the current exams with adjustments for implementing new and different development processes. The start-up time required to orient subject matter experts to ECD and to the iterative nature of the work added strain to the project timeline and made the work more resource-intensive than originally expected. The demand for communication and discussion time was very high. Consequently, lack of time to complete work during face-to-face meetings was frequently an issue. As a result, 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 and presented to the group for discussion, revision, and eventual endorsement. The SMEs sometimes viewed the iterations as having to do things over, thinking it must have been initially done the wrong way. The process has now been completed with a few exams, so it is possible to better identify the amount of time needed for stages of the process and to set expectations with stakeholders about when iteration is a natural part of the process and should be built into the project plan.

While these kinds of iteration are expected in a design process, sufficient time should be allotted for gathering all information, and SMEs should be reassured that iterations based on more information do not indicate that previous iterations (based on less information) were in error. There are several ways to mitigate and manage the issues associated with iteration:

- Make decisions ahead of time about any validation or sign-off steps that will need to be incorporated into the process, as well as the time to address concerns from those steps.
- Emphasize the iterative nature of the work from the beginning with all stakeholders involved in the process.
- Build sufficient and additional time—more time than you think it will actually need—into the project plan for gathering and incorporating all needed information, as well as time for iteration.
- For those points where you know that iteration is likely, run "checks" ahead of the final product being completed, as much as possible. For example, as the domain model is written, have the same subject matter experts who write the claims and

evidence also try to develop ALDs and at least a few task models for a small sample of claims and evidence. This should occur early in the process before all of the claims and evidence are drafted. This review allows lessons learned to be incorporated into the claim and evidence writing efforts.

- Create a checklist for each stage of the process that helps to identify if/when all decisions are "complete" and/or if iteration is necessary. See Table 2 for an example.

Table 2.

Iteration checklist

Domain Analysis	**Yes**	**No**
1. Are all Big Ideas at a similar and appropriate level of specificity?		
2. Are all Enduring and Supporting Understandings at a similar and appropriate level of specificity?		
3. Is there consensus that all appropriate Big Ideas are included?		
4. Is there consensus that all appropriate Enduring and Supporting Understandings are included?		
5. Is there consensus that all of the appropriate reasoning processes or practices that students use when they interact with content are included?		
6. Is external validation of the domain analysis required? If so, has it been completed?		
Domain Model		
1. Are there claims written to all Big Ideas?		
2. Are there claims written to all Enduring and Supporting Understandings?		
3. Are there claims written to all practices (i.e., skills)?		
4. Does the level of the specificity of the claims and evidence align to the purpose of the test?		
5. Are the claims at the appropriate level of specificity to support evidence statements that are more than a restatement of the claim?		
6. Are the claims and/or evidence statements able to support generation of task models?		
7. Is external validation of the domain model required? If so, has it been completed?		
Assessment Framework		
1. Do all task models meet the evaluation criteria (e.g., see Table 1)?		
2. Can items be easily mapped to claims and/or evidence statements and ALDs?		
3a. Is there additional information outside of the claims, evidence, ALDs, and task models that is necessary to determine test specifications, such as the desired distribution of claims, tasks, ALDs on the assessment?		
3b. If yes, may this additional information reveal necessary changes to the domain analysis or model?		
4. Is external validation of any part of the assessment framework required? If so, has it been completed?		

Conclusion

Implementation of ECD for test development is challenging work. The potential benefits are assessments that better reflect and measure what is taught and valued in the classroom, and resulting score inferences that are strongly supported by an evidentiary argument. Such benefits make the time and effort worthwhile and justify the increased resources required at the beginning of the design endeavor. We hypothesize, also, that the use of ECD will become less resource-intensive once it is employed more broadly in assessment design and development, and those individuals who are working with ECD in research and in practical applications of the methodology identify pathways through the major challenges.

No assessment is created without constraints and all resources must be justified. The argument that eventually efficiencies will be realized through the use of task models can be less persuasive than expected because the item development costs for most assessment programs are, proportionately, very small compared to costs of administration and scoring. Consequently, justifying the resources required for ECD falls largely on the argument that the benefits include an improved validity argument and, as a component of validity, improved score comparability. Although this argument can be arcane and far-reaching to those who make decisions about resources, it can be compelling when discussed in light of the consequences of producing scores that are not comparable or are not as predictive of the intended criterion as desired.

Many fields – science, medicine, law, manufacturing quality control – have a long history of or have evolved toward an evidence-based approach to investigation. Evidence-centered assessment design posits explicit evidentiary standards at the beginning of design rather than post hoc. Although this approach may be relatively novel to the field of measurement, ECD can be seen as a product of its time, where the concept of evidence is manifest everywhere.

In this article, we have identified the most challenging areas that we have encountered in implementing ECD for several AP exams. By no means are these challenges insurmountable, but they do require more time, thought, and research than traditional test development. We hope that our descriptions of the challenges and potential mitigation strategies are useful to those already implementing ECD and encouraging to those considering implementation of ECD for their own testing programs.

References

College Board. (2011). AP French Language and Culture Course and Exam Description. http://apcentral.collegeboard.com/apc/public/repository/AP_FrenchLangCED_Effective_Fall_2011.pdf

College Board. (2012). AP Biology Course and Exam Description. http://apcentral.collegeboard.com/apc/public/repository/AP_BiologyCED_Effective_Fall_2012_lkd.pdf

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global E-learning program. *The International Journal of Testing*, *4*, 295-301.

Ewing, M., Packman, S., Hamen, C., & Thurber, A. (2010). Representing targets of measurement within evidence-centered design. *Applied Measurement in Education, 23*(4), 325-341.

Hendrickson, A., Huff, K., & Luecht, R. (2010). Claims, evidence, and achievement-level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education, 23*(4), 358-377.

Huff, K., Alves, C., Pellegrino, J., & Kaliski P. (in press). Using evidence centered design task models in automatic item generation. In M. Gierl & T. Haladyna (Eds.), *Automatic item generation*. New York, NY: Informa UK Limited.

Huff, K., & Plake, B. S. (2010). Innovations in setting performance standards for K-12 test-based accountability. *Measurement: Interdisciplinary Research & Perspective, 8*(2), 130-144.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*(4), 310-324.

Kaliski, P., Huff, K., & Barry, C. (2011, April). *Aligning items and achievement levels: A study comparing expert judgments*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25*, 6–20.

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

National Research Council. (2000). *How people learn: Mind, brain, experience and school*. 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.

The Partnership for Assessment of Readiness for College and Careers (PARCC). (2012, June). *PARCC Assessments in the Making: A Principled Assessment Design Approach*. 2012 National Conference on Student Assessment, Minneapolis, MN.

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

Schneider, M. C., Huff, K. L., Egan, K. L, Tully, M., & Ferrara, S. (2010, May). *Aligning achievement level descriptors to mapped item demands to enhance valid interpretations of scale scores and inform item development*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.