**Evidence-Centered Design**

Of utmost concern when developing an assessment is the validity of the scores produced by the assessment, and the decisions based on those scores. Traditionally, gathering of validity evidence begins after the assessment has already been developed. The problem with this approach is that items are not built specifically to support the inferences that the assessment intends to make (Mislevy, Almond, & Lukas, 2003; Pellegrino, 2014). Thus, wonderfully complex items can be written that provide rich data, but if it is unclear what to do with that data, or if it is not clear how the data support the use of the assessment, then time and money has been wasted. However, with the advancement in technology that has made these more complex items possible, have also been advancements in assessment engineering (Gierl, Zhou, & Alves, 2008) and evidentiary reasoning (Schum, 1994). Out of later these later advancements comes a solution to the problem of collecting validity evidence after the assessment is already built: evidence-centered design (Mislevy, 2007).

Evidence-centered design is a framework for assessment development that is focused on building validity evidence into the assessment from the conceptual to inferential elements. In short, evidence-centered design is premised on asserting the inferences one wishes to make and the evidence required to make those inferences prior to the development of the assessment (Bond, 2014). Tasks and items can then be written in a way to provide the needed evidence. Thus, the assessment is built in such a way that the precise evidence needed from test takers to make inferences about their underlying ability levels is evoked. This framework has been used in both standard assessment forms, simulation assessments (Mislevy, 2013), and game based assessments (Rupp, Gushta, Mislevy, & Shaffer, 2010). This paper seeks to outline the evidence-centered design (ECD) framework by exploring the most widely cited and used ECD approach that was made popular by Mislevy and his colleagues (Mislevy, Behrens, Dicerbo, & Levy, 2012; Mislevy & Haertel, 2006; Mislevy & Riconscente, 2005). This paper will then explore some challenges associated with implementing the ECD framework, along with critiques and modifications that others have made to Mislevy’s original framework.

**The Evidence-Centered Design Framework**

The ECD framework consists of five layers, as laid out by Mislevy, Steinberg, and Almond (2003). These layers are domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery. ECD is inherently an argument-based approach to providing validity evidence. Thus, the layers build upon each other, from defining the construct of interest to reporting scores, in a way that directly links the observable data to the claims that the assessment makes. These links form the validity argument. To further explore these links, each layer will be explained individually.

**Domain Analysis.** The domain analysis is primarily concerned with gathering information about and defining the construct that is to be assessed. In essence, this layer is concern with defining the construct, a step that is vital to any assessment process (Messick, 1994). This includes examining how knowledge of the construct is learned, utilized, and communicated (Mislevy & Haertel, 2006). This is where terminology specific to the construct should be defined, along with situations where knowledge of the construct is used and any interactions that the construct may have with other similar or related constructs (Mislevy & Riconscente, 2005). It is important to denote situations in which knowledge of the construct is valued, as these situations will become the basis for assessment tasks.

**Domain Modeling.** The domain modeling layer takes the information that was gathered during the domain analysis, and organizes it into a narrative argument (Mislevy & Haertel, 2006). This process begins with the *claims*. A *claim* is a statement made about an individual’s proficiency in a construct. In other words, what is the assessment trying to say? There may multiple claims for a single assessment, or just one claim (competent/incompetent, or pass/fail). The next piece of the argument is the *data.* The *data* are the features that are important for the tasks to include. This should include goals for the tasks and possible stimulus materials. The *data* will lead to design patterns, which are generally centered on specific knowledge, skills, and abilities that are relevant to the construct. Design patterns serve as a sort of template for tasks that can be used to evoke evidence of the knowledge, skills, or abilities represented (Ewing, Packman, Hamen, & Thurber, 2010). This part of the argument would also contain potential products that an individual might produce (e.g., say or do something). This will also aid in developing tasks in later layers. The final part of the argument is the *warrants*. A *warrant* is a logical reasoning connection between the *data* and the *claim*. The *warrant* answers the question, “Why does this *data* represent evidence for the *claim* I am trying to make?” In this way the validity can be seen getting built into the assessment. The *claims* of the assessment have been specified, the types of *data* that the assessment will produce have been described, and the *warrants* that support the connection between the *claims* and *data* have been reasoned out for others to judge or critique.

**Conceptual Assessment Framework.** The next layer, the conceptual assessment framework, conveys the argument that was laid out in the domain modeling as actual tasks for the assessment. However, the conceptual assessment framework is not limited solely to the items of the assessment (Velasquez & Tokac, 2014). This layer also involves the creation of test specifications, evaluation procedures, and the calibration of measurement models that will be used to score the assessment. Within the conceptual assessment framework are three models that describe the features of the assessment: the *student model*, *task models,* and *evidence models*.

The *student model* represents the person-level variables an assessment is attempting to measure. This can be proficiency on a construct to determine licensure qualifications or the *student model* can be multidimensional, such as a test that aims to provide detailed feedback on several areas rather than one overall score (e.g., a math assessment that provides feedback on multiplication and division separately, rather than a single math score). The *task models* describe how the individuals taking the assessment will provide the evidence that was specified in the domain analysis. *Task* *models* describe what the test items or work products will look, as well as what stimulus materials may be needed (Luecht, 2009). For any given assessment, there will be many different *task models*, each associated with a different piece of evidence, or set of knowledge, skills, and abilities. Thus, the actual number of *task models* will be dependent on how much and what types of evidence are needed, as described in the domain modeling. Finally, the *evidence model* connects the *task model* back to the *student model*. The first step of the evidence model is the evaluation of the *task model*. How will the *task models* be scored? This can included automated scoring procedures or rubrics and training for human scoring (Mislevy, 2010). The type of evaluation process that is utilized will be largely dependent on the type of evidence that the *task model* is attempting to provide. The final step of the *evidence model* is synthesizing the evaluated *task models* into a measurement model (Stout, 2002), such as an Item Response Theory model.

**Assessment Implementation.** The assessment implementation layer is where the assessment actually gets built. This involves developing all of the operational pieces of the assessment, as specified by the conceptual assessment framework (Rupp, Nugent, & Nelson, 2012). Included in this layer is writing and reviewing items, finalizing and quality checking scoring procedures (both automated and manual), collecting pilot data to estimate the parameters of the chosen measurement model, and assembling the final fixed form assessment, or algorithm for adaptive tests. Theses steps are all common to assessment development. However, with the ECD framework, these steps are all now linked back directly to the validity argument.

**Assessment Delivery.** The final layer of ECD is the assessment delivery. This involves individuals actually taking the assessment and receiving scores and feedback (Mislevy, Steinberg, Almond, & Lukas, 2006). Thus, this is where score reports would be developed, keeping in mind the claims that were laid out in the domain modeling. In the Mislevy model (Mislevy & Haertel, 2006), this layer has four distinct processes. The first is the *activity selection process* where tasks are selected from a “task library” (or item bank). This could also be the selection of which form of an assessment an individual will receive. Following the task selection is the *presentation process*. As the name implies, this describes how the task will be presented and how the result, or product, of the task will be recorded. These results then become a part of the *evidence identification process*. This is the scoring of the task, following the procedures and methods that were specified in the *evidence model* of the conceptual assessment framework. Finally, these item scores are passed to the *evidence accumulation process*, which employs the *measurement model* specified in the conceptual assessment framework to produce a test-level score, as well as a score report with feedback for the individual.

**Challenges to Implementing the ECD Framework**

The main advantage of ECD is the built in validity argument (Brennan, 2010). ECD takes an evidence-claim approach to creating a validity argument that is very similar to the work of Kane (2006), who lays out validation evidence in the terms of an interpretation-use argument. Thus, when an assessment is built using the ECD framework, all the evidence needed to support the use of test scores to make a decision is already completed. Additionally, the *task model* approach is helpful to item writers, as they can help focus the tasks on the specific knowledge, skills, or abilities that are needed to provide the required evidence (Bauer, 2014).

However, ECD is not without its drawbacks. The first is that ECD can be difficult for assessment designers to learn due to the jargon that is used within the ECD framework (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004). The jargon can be a barrier to understanding what different aspects of ECD are referring to, even though many of elements of ECD are common to standard assessment design. Behrens et al. (2004) notes that although it is possible to translate ECD terminology into more familiar concepts, this is undesirable as using standard terminology is likely to lead to the use of more standard assessment methods, rather than following precisely what ECD prescribes. Thus, it has been advocated that before implementing an ECD framework, that extensive training may be required, as simply reading articles about ECD unlikely to provide the full understanding that is necessary in order to apply this framework correctly, and receive the full benefits (Zieky, 2014).

One of the largest implementations of ECD occurred when the Advanced Placement assessments (administered by the College Board) were redesigned (Huff, Steinberg, & Matts, 2010). Although the overall implementation was successful, the test developers noted several challenges that they encountered during the process. First, they note that it can be difficult to incorporate theories of how individuals learn into the assessment design. How knowledge is acquired is specified in the domain analysis and is a critical part of construct (Mislevy, 2014), however, this aspect receives little attention in the later layers. They note that learning theory can be incorporated within the ECD, and it has been suggested that learning theory be used to inform performance level descriptors (Plake, Huff, & Reshetar, 2010), which can then in turn be incorporated into the *task models* (Hendrickson, Huff, & Luecht, 2010).

Additionally, they note that in the ECD framework it can be difficult to get the right amount of specificity when stating the *claims* in the domain modeling (Hendrickson, Ewing, Kaliski, & Huff, 2013). They suggest considering *claims* in the context of the data that could be generated. A *claim* that is too broad will result in huge amounts of data, as more knowledge, skills, and abilities will be included with ever-broader *claims*. Conversely, if a *claim* is too narrow, there will be very few *task models* that can be written, resulting in too few data. Thus, when developing the *claims*, it is important to consider the data that was described in the domain analysis. Which data would a *claim* at this grain size allow the *task models* to assess? Thus, the ECD framework begins to become an iterative process (Mislevy, 2011). Indeed, Hendrickson et al. (2013) advocate for an iterative approach to ECD, allowing for changes to be made to the assessment design as needed.

The third challenge they note is the difficulty in developing the *task models*. Much like the *claims*, it is important to get the *task models* at the right grain size. If the *task model* is too broad, then items may be variable in the type of evidence they are provided, however a too narrow *task model* doesn’t allow for an adequate number of unique items. Two strategies were used to combat this challenge. The first option was to have subject matter experts break a *claim* down into a series of steps, and each step could then be the focus of a task model. An alternate option is to develop criteria that can be used to evaluate the *task models.* In this option, we again see iteration, where *task models* can receive feedback and be reworked to better reflect the argument laid out in the domain modeling.

The final challenge noted with regards to the Advanced Placement test was directly related to the iterative ECD approach (Ewing et al., 2010). That is, there was difficulty in determining where the iteration should occur (Hendrickson et al., 2013). Although iteration can be built into the process, it was common for unplanned iteration to become necessary based on team discussions, and this added additional time and cost to the development of the assessment. To counteract this, the authors advocate for iteration within each layer (e.g., domain analysis, domain modeling, etc.) to ensure consistency of the elements of the layer. Similarly, they suggest that when moving between layers, that previous decisions that were made between other layers also be revisited. This is to ensure that the assessment remains on a consistent path, focused on the overall validity argument. Finally they suggest have a stakeholder review of all the products produced by the ECD framework. This allows for yet another chance to iterate and make changes to ensure that the stakeholders are receiving the best possible assessment to fits their needs.

**Summary**

ECD offers an alternative framework to the standard procedures for assessment design. This framework is centered around and argument approach to validity evidence, with the argument built directly into the assessment. This is a significant improvement over standard procedures, which may not accurately represent the construct fully or have clear connections between the data that is being gathered and the claims the that assessment aims to make. This improvement is not without costs, however. There are many challenges to implementing the ECD framework in a large-scale operational setting, but by learning from the experiences of previous implementers of ECD, it is possible to work through the ECD process in a more efficient manner.

**References**

Bauer, M. (2014). The process of designing task features. *Measurement, 12*(1-2), 34-36. doi: 10.1080/15366367.2014.922367

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

Bond, L. (2014). A breif note on evidence-centered design as a mechanism for assessment development and evaluation. *Measurement, 12*(1-2), 37-38. doi: 10.1080/15366367.2014.921486

Brennan, R. L. (2010). Evidence-centered assessment design and the Advanced Placement Programs: A psychometrician's perspective. *Applied Measurement in Education, 23*(4), 392-400. doi: 10.1080/08957347.2010.510973

Ewing, M., Packman, S., Hamen, C., & Thurber, A. C. (2010). Representing targets of measurement within evidence-centered design. *Applied Measurement in Education, 23*(4), 325-341. doi: 10.1080/08957347.2010.510959

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *The Journal of Technology, Learning, and Assessment, 7*(2).

Hendrickson, A., Ewing, M., Kaliski, P., & Huff, K. (2013). Evidence-centered design: Recommendations for implementation and practice. *Journal of Applied Testing Technology, 14*.

Hendrickson, A., Huff, K., & Luecht, R. (2010). Claims, evidence, and acheivement-level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education, 23*(4), 358-377. doi: 10.1080/08957347.2010.510966

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. doi: 10.1080/08957347.2010.510956

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.

Luecht, R. M. (2009). *Adaptive computer-based tasks under and assessment engineering paradigm*. Paper presented at the 2009 GMAC Conference on Computerized Adaptive Testing.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469. doi: 10.3102/0013189x07311660

Mislevy, R. J. (2010). Design under constraints: The case of large-scale assessment systems. *Measurement, 8*, 199-203.

Mislevy, R. J. (2011). Evidence-centered design for simulation-based assessment. *CRESST Report 800*.

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Mil Med, 178*(10 Suppl), 107-114. doi: 10.7205/MILMED-D-13-00213

Mislevy, R. J. (2014). Postmodern Test Theory. *Teachers College Board, 116*.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Reserach Report, RR-03-16*.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining, 4*(2), 12-48.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology*: SRI International.

Mislevy, R. J., Steinberg, L., Almond, R. G., & Lukas, J. F. (2006). Concept, terminology, and basic models of evidence-centered design. In D. M. Williamson, I. I. Bejar & R. J. Mislevy (Eds.), *Automated Scoring of Complex Tasks in Computer-Scoring* (pp. 15-47). Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*(2), 65-77. doi: 10.1016/j.pse.2014.11.002

Plake, B. S., Huff, K., & Reshetar, R. (2010). Evidence-centered assessment design as a foundation for achievement-level descriptor development and for standard setting. *Applied Measurement in Education, 23*(4), 342-357. doi: 10.1080/08957347.2010.510964

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning, and Assessment, 8*(4).

Rupp, A. A., Nugent, R., & Nelson, B. (2012). Evidence-centered design for diagnostic assessment within digital learning environments: Integrating modern psychometrics and educational data mining. *Journal of Educational Data Mining, 4*(1), Article 1.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. New York, NY: Wiley.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*(4), 485-518.

Velasquez, G., & Tokac, U. (2014). Using evidence-centered design to diagnose proficiency in solving story problems. *Journal of Multidisciplinary Research, 6*(2), 65-76.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa, 20*(2), 79-87. doi: 10.1016/j.pse.2014.11.003