Principled Assessment Design for the Performance Assessment of Competency Education (PACE)

Scott Marion

Draft: August 2, 2017

Introduction and Rationale

How should we design performance-based assessments to support learning, instructional, and accountability purposes? The performance assessments used to evaluate student learning of key competencies in PACE are well-suited to using a principled approach to design such as Evidence Centered Design (ECD; Mislevy, 1994, 1996) or following the assessment triangle as articulated in *Knowing What Students Know* (Pellegrino, Chudowsky, and Glaser, 2001). Principled design is an attempt to move from "one-off" designs to more replicable task designs and templates. It is also an effort to design for validity from the initial conceptualization of the task instead of a post-hoc analysis. Principled assessment design requires task developers to be clear about:

- What makes a task harder or easier?
- What makes a task more or less complex?
- What makes a task more or less likely to elicit evidence of student achievement of the key learning targets?

These questions are usually thought of implicitly, if at all, in task design, but current work using principled assessment design such as with the Advanced Placement program and with the consortium assessments (i.e., PARCC, Smarter Balanced, and NCSC) has demonstrated the practical and theoretical advantages of answering such questions explicitly.

Importantly, principled assessment design intends to ensure that assessments are based on research-based models of learning. Bob Mislevy, the originator of Evidence Centered Design, once famously noted "It is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of 20th century statistics to 19th century psychology (Mislevy, 1993, p. 19)." Adherence to outdated, naïve, and/or implicit notions of learning is an impediment to the design of performance assessments of deeper learning as well as to the usefulness of such assessments for improving learning and instruction. Principled assessment design is an attempt to ensure that assessments are built on modern theories of learning to provide a more robust framework for the design and interpretation of assessment results.

Too often assessments are designed by simply trying to match test questions or tasks to individual standards or even competencies, but this leaves us wanting in how to meaningfully interpret the results. We want information about the degree to which students are developing competence in the domain, but unless the assessment is purposefully designed to provide such information, we are likely to just get an estimate of general achievement.

Principled Assessment Design

Bob Mislevy and his colleagues (e.g., 2003, 2006) proposed Evidence Centered Design as a very complex test design and interpretation framework for better evaluating and supporting inferences derived from test scores. In 2001, the National Research Council published *Knowing What Students Know*: *The Science and Design of Educational Assessment* (Pellegrino, Chudowsky, & Glaser, 2001), which synthesized a tremendous body of learning and measurement research and set an ambitious direction for the development of more valid assessments. *Knowing What Students Know* (KWSK) built off of Mislevy's (1996) notion of assessment as a process of reasoning from evidence and previous NRC work synthesizing research on human learning (Bransford, Brown, and Cocking, 2000). The authors of *Knowing What Students Know* used the heuristic of an "assessment triangle" to illustrate the relationship among learning models, assessment methods, and inferences from assessment scores. We provide a little detail here because it serves as an important background to understanding ECD.

Cognition refers to the empirically-based theories and beliefs about how humans represent information and develop competence in a particular academic domain (Pellegrino et al., 2001). The theories of "learning and knowing" that help explain varying levels of performance in a particular domain are crucial for the design and interpretation of assessments. The observation vertex of the triangle refers to “a set of specifications for assessment tasks that will elicit illuminating responses from students” (Pellegrino et al., 2001 p. 42). The design of items or tasks is based upon the belief that those particular assessment events will allow students to demonstrate their understanding in a given domain, based upon a particular view of learning and knowing. The interpretation component in this diagram includes all of the methods and analytic tools used to make sense of and reason from the assessment observations (Pellegrino et al., 2001).

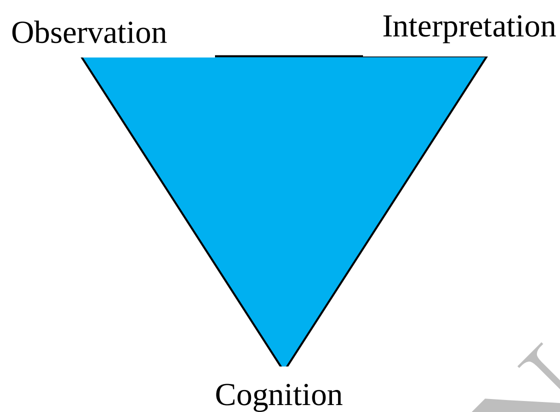

Figure 1. The Assessment Triangle (from NRC, 2001, p. 39)

Evidence Centered Design

While the assessment triangle offers many advantages, we have found that for our purposes, using the basics of ECD provides an understandable and powerful framework for helping educators design high quality performance tasks. In its simplest formulation, the core of the ECD framework comprises a student model, an evidence model, and a task model (see Figure 2 below). The student model describes the intended construct(s) or learning outcome(s), the evidence model, which links the task and student models, describes the nature of the evidence that would convince one that the students mastered the intended knowledge and skills associated with the construct, and the task model describes the types of assessments that will elicit the desired evidence (see Haertel, et al., 2016).

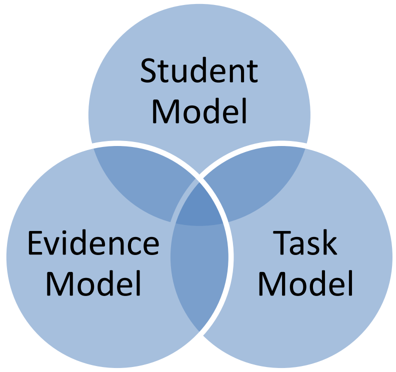

Figure 2. Schematic of the basic ECD model.

The Student Model

The student model is analogous to the cognition vertex in the assessment triangle but focuses on the claims that we intend to make and support based on this learning and assessment experience. In defining the student model, assessment designers are asked to specify exactly what they want students to know and how well do they want them to know it. This requires a very careful examination of the "construct" to unpack the thing we want students to know and how well we want them to know it. The "construct" is not just a content standard or even set of content standards or competencies. Rather, construct refers to a hypothesized attribute such as reading comprehension or scientific inquiry that is based on a theoretical understanding of how various knowledge, skills, and dispositions come together to define this attribute as well as how learners progress in their mastery of this construct along a continuum from fragile to deeper understanding.

Evidence Model

The evidence model is really like a thought experiment where one needs to describe what sort of **evidence** would convince us that the student demonstrated the knowledge and skills described in the student model. For example, if the student model focused on the construct of argumentative writing, an evidence model might include such expectations as high-quality performance on a series of diverse pieces of argumentative essays on a range of topics. Ultimately, we need to ask ourselves, what will we accept as evidence that the student has mastered the knowledge and skills that define the student model (construct)?

The evidence model is almost always bypassed in task design in the rush to create items and tasks. In order to avoid a tail wagging the dog phenomenon, specifying the desired evidence *a priori* will help ensure that the focus is on the construct and not simply on the assessment tasks.

<u>Task Model</u>

Once the evidence model is specified, we can then turn our attention to task design. Notice that we do not start with the tasks and try to retrofit the learning goal. The task model requires designers to consider the nature of the tasks that students will perform to demonstrate and communicate their knowledge. In the case of PACE, we use a task design template to ensure that performance tasks are designed to best represent the intended learning targets.

<u>An Example</u>

The following example from the Advanced Placement program (Huff & Plake, 2010) helps to highlight what is required to specify the student model. Note that the enduring understanding represents the major claim in that the designers would like to have evidence to support the claim that students are able to demonstrate an understanding that chemical reactions are represented by a balanced chemical reaction that identifies the ratios with which reactants react and products form. As shown below, the big idea and enduring understanding provide grounding in the major ideas of the domain, but the supporting understandings help provide the level of detail necessary to support evidence and task conceptualizations.

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

Enduring Understanding: Chemical reactions 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 the 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.

Figure 3. From Huff & Plake (2010). An example content outline in chemistry for one big idea.

These major concepts and understandings are then further delineated into the component knowledge and skills that support these big picture understandings.

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

Figure 4. From Huff & Plake (2010). Defining knowledge and skills related to the big idea.

Universal Design for Learning

The use of principled assessment design has tremendous advantages for the design of assessments, including the types of curriculum-embedded performance tasks used in PACE and similar projects. But what about students with disabilities, English learners, or others struggling to access the content in expected ways.

Universal Design for Learning (UDL) is an educational framework based on research in the learning sciences that guides the development of flexible learning environments that can accommodate individual learning differences. The UDL framework, first defined by David H. Rose and the Center for Applied Special Technology (CAST) in the 1990s, calls for creating curriculum from the outset that provides:

- *Multiple means of representation* to give learners various ways of acquiring information and knowledge,
- *Multiple means of expression* to provide learners alternatives for demonstrating what they know, and
- *Multiple means of engagement* to tap into learners' interests, challenge them appropriately, and motivate them to learn

UDL has been applied to assessment design increasingly over the past 15 years or so. In fact, when asked about the relationship of UDL to principled assessment design, Mislevy responded:

> *UDL prompts you to target learning goals; you identify what we call the "focal knowledge, skills, and abilities" or "focal KSAs," that you want your students to develop. When applying UDL to assessment, you are evaluating these focal KSAs in order to determine if students are making progress in those capabilities. UDL also encourages us to carefully consider all of the knowledge, skills, or abilities that might tangentially be involved in assessing the focal ones. These "non-focal KSAs" might prevent students from accurately being able to demonstrate what they know and what they can do. For example, students with a visual impairment might do poorly on a science assessment not because they do not know the content*

but because they are unable to see the material. Other students may do poorly on a specific item simply because they were not given some construct-irrelevant information that they would need to know in order to interact with the task. In both of these examples, non-focal KSAs interfere with students' learning and performance on tests, and lead to invalid assessment. UDL pushes us to think about the ways in which we can support students' non-focal KSAs so that we can target and address the actual learning goals (p.7).

This applies to our work of performance assessment design throughout the design and implementation stages. Instead of trying to "fix" or accommodate tasks after the fact, UDL directs us to intentionally design tasks for the widest range of student needs possible. For example, we should avoid:

- Using extraneous words that potential distract students from the main learning target of the task
- Using idioms or culturally-specific language
- Crowding text and/or graphics too closely on the page
- Using graphics that require certain levels of visual acuity to understand

Summary

This a working document. We will develop and share grade- and subject-specific examples in coming months and we will be updating the task template to better fit the principled assessment design processes outlined here. While some of the steps outlined in this document may appear more cumbersome compared to just designing a task, I argue that following the actions outlined in this document will lead to significantly higher quality tasks than those developed in a more ad-hoc manner. Importantly, a principled design process will improve the efficiency and replicability of our task design efforts.

References

Bransford, Brown, & Cocking (Eds.). (1999). How People Learn: Brain, Mind, Experience, and School. National Research Council (in the process of being updated).

Gordon, D. T., Gravel, J.W., & Schifter, L.A. (2011). Perspectives on UDL and Assessment: An Interview with Robert Mislevy. In Gordon, D.T., Gravel, J.W., & Schifter, L.A. (2009). *A policy reader in universal design for learning.* (pp. 209-218). Cambridge, MA: Harvard Education Press.

Haertel, G.D., Vendlinski, D. R., DeBarger, A., Cheng, B.H., Snow, E.B., D'Angelo, C., Harris, C.J., Yarnall, L., & Ructtinger, L. (2016). General introduction to evidence-centered design. In Braun, H. (ed). *Meeting the Challenges to Measurement in an Era of Accountability*. Pp. 107-148. New York, NY: Routledge, Taylor & Francis Group.

Huff, K. & Plake, B. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*, 307-309.

Mislevy, R. J. (1993). Foundations of a new test theory. In Frederiksen, N., Mislevy, R. J., and Bejar, I. I. (Eds.), *Test theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J. (1996). Evidence and inference in educational assessment. CRESST Technical Report No. 414. https://pdfs.semanticscholar.org/5eae/5388283e95a3a8f3e5b291bcae7f3558dd44.pdf

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

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

Pellegrino, J., Chudowsky, N. & Glaser, R. (2001). *Knowing What Students Know*: *The Science and Design of Educational Assessment.* National Research Council of the National Academy of Sciences. www.nap.edu

Rose, D.H, & Meyer, A. (2002) *Teaching Every Student in the Digital Age: Universal Design for Learning*Alexandria, VA: ASCD.