Validating Assessment for Learning: Consequential Systems Approaches

James Patric Van Haneghan
University of South Alabama

ABSTRACT: In order to adequately evaluate assessment for learning, expanded approaches to validity need to be considered. The purpose of this manuscript is to explore what is necessary to evaluate claims that assessment facilitates learning. Messick's (1994) concept of consequential validity provides one lens for determining the learning consequences of assessment. His approach suggests that learning consequences are a special case of the general concept of consequential validity and should be evaluated from that perspective. The systems approach developed by Frederiksen and Collins (1989) provides another perspective of how assessments can be designed with learning consequences in mind. Their model provides a transparent way to link assessments to learning by making teaching to the test a valid activity. The adequacy of both models for evaluating claims of learning from assessment is explored. Based on the analysis, a model for evaluating the validity of evidence for assessment for learning is outlined.

KEYWORDS: *consequential validity, educational evaluation, formative assessment, learning, performance assessment, systemic validity,*

Over the last two decades, there has been an accumulation of evidence that assessment can provide feedback to students and teachers to help them to facilitate better learning (e.g., see Black & Wiliam, 1998; Pellegrino, Chudowsky, & Glaser, 2001). The evidence was so compelling to policymakers in the United Kingdom, that assessment for learning has become an important initiative that has been a focus for schools in that country (e.g., see Mansell, 2008). Interestingly, there has been concern about the authenticity of the implementation (Ward, 2008). The *Times Educational Supplement* quotes Paul Black, a leading scholar on assessment for learning, as stating that "This [the program in the UK] is not assessment for learning. It may help learning, but it is not what I and colleagues have been writing about and helping teachers with since 1998" (Mansell, 2008, p. 7). Such criticism suggests that the validity of assessment for learning approaches is important. As has been noted by many authors (e.g., Shepard, 2006) most approaches to validity have focused on summative assessments using tests. Approaches to validation of formative assessments and more performance-oriented assessments has lagged behind somewhat. In order to adequately evaluate assessment for learning, expanded approaches to validity need to be considered.

The purpose of this manuscript is to explore what is necessary to evaluate claims that assessment facilitates learning. First an analysis of what assessment for learning actually reflects

is presented. Second, an analysis of two validity constructs surrounding assessment for learning will be undertaken. One is Messick's (1994) concept of consequential validity, which provides one lens for determining the learning consequences of assessment. His approach suggests that learning consequences are a special case of the general concept of consequential validity and should be evaluated from that perspective. The second approach is the systems approach developed by Frederiksen and Collins (1989). Their model provides a transparent way to link assessments to learning by making teaching to the test a valid activity. Based on the analysis of these two models and more recent work, I will develop a more complete model of assessment for learning validity. Practical implications of the model for evaluating assessment for learning systems will then be described.

What is Assessment for Learning?

The concept of assessment for learning is complex because undoubtedly almost any assessment can provide some information about learning to someone. What most individuals who study assessment for learning refer to is the use of formative assessment in classrooms that helps improve student learning (Black & Wiliam, 1998). However, even summative assessments provide some information that might facilitate or inhibit learning. For example, students who do poorly on an exam might learn from reflecting on their preparation that they did not study enough. Students who do poorly, and attribute their performance to a lack of ability in an area, may confirm their belief from the test information. Well-designed exams that adequately sample aspects of a learning domain can provide feedback about areas of learning that can be improved.

A further issue that makes it more complex is that assessment for learning benefits may operate in different ways. Students' learning may be enhanced directly from their reflection on an assessment. Students' learning may be enhanced by the reflection of a teacher that leads to changes in instruction that better supports learning. Or, learning can be enhanced through some combination of teacher and student insights. In situations where students are working on intelligent tutoring systems, tailored instruction that emerges from online assessments of how students approach tasks provides the impetus for learning. Yet, in other situations, students work together on assessment and evaluation of their work. Hence, assessment for learning can happen through the assessment of a peer or more able student in conjunction with assessment artifacts (score sheets, rubrics, feedback sheets, performance outcomes, and so on).

Pellegrino et al. (2001) talk about an assessment triangle that includes three interrelated components: the theory of cognition and learning for the task, the observation of the student, and interpretation of the assessment. Along with the triangle, their report talks about the importance of the reasoning process about the evidence produced by an assessment. Hence, in the interpretation of assessments for learning, the reasoning processes of the teacher, the reasoning processes of a student or students, and the joint reasoning of students and teachers together often need to be examined. Further, these reasoning processes may take place with assessment artifacts or tools associated with the particular assessment. For example, a dialogue between a student and a teacher may take place surrounding a student's level of performance on a rubric and how it can be improved. Thus, the interpretation and reasoning that takes place surrounding a formative assessment creates a dialogue that may include the student, the interpretative evidence, and the teacher. However, the process of reasoning may involve a self-assessment of a student engaging in an analysis of an assessment artifact, students reasoning together about of an assessment artifact, or many teachers and students

reasoning around an assessment artifact. Or, it might not involve the decisions of a computer-based tutor that uses assessment information to tailor instruction to enhance learning.

The different varieties of interactions and reasoning processes that surround assessment for learning mean that validating assessment for learning requires a model that takes into account the different ways in which assessment for learning takes place. Thus, a valid model will have to consider the nature of the assessment artifacts, the model of reasoning or decision making around assessment artifacts, and whether those engaged in reasoning with an assessment comprehend it in ways that can help students learn. Further, the model needs to take into account how more automated feedback from assessment in the context of computer-based intelligent tutoring systems can be used as effective tools by students and teachers to help guide learning.

Systems Approach to Validity

Frederiksen and Collins (1989, 1996) developed a model that provides a starting point for creating a model of validity that can be used in assessment for learning situations. The key element of the model is that assessment is set up to be direct, systematic, and comprehensible by instructors and students. It starts with assessments that reflect meaningful real tasks rather than indirect measures like multiple choice tests that correlate with meaningful tasks but do not reflect meaningful performances. The approach is systematic in that task performance is scored using rubrics that focus on the set of traits that make up the characteristics of successful performance. These characteristics are defined so that students and teachers both understand these rubrics so that the dialogue around task performance can lead to improved performance.

The validity of assessments using this model appears at first blush to be very straightforward. One should be able to collect evidence that the task and scoring rubric are valid (perhaps through examining the judgments of experts). One should be able to determine that the dialogue between students and teachers surrounding the rubric scores is facilitative of learning. And, when asked, all parties involved in the assessment process should be able to talk in meaningful ways about the assessment. Finally, given that the rubric is clear, learning consequences that will improve task performance can be clearly specified. In many ways this model works well in defining situations where assessment for learning can take place.

Messick's Critique and Model for Assessment for Learning

On the other hand, Messick (1994, 1996) provided evidence that complicates the process of validation of performance assessments like the ones developed within the context of Frederiksen and Collins model. The crux of his critique focuses on the claims about task validity. In particular, Messick noted that the tasks often used in performance-based assessments were open to criticism from the point of view of construct validity. He points out that claims about tasks being "authentic" and "direct" are claims about the validity of those tasks. He noted that specific tasks might underrepresent the domain of interest. For instance, if one uses a problem-solving task as an assessment, then the specific knowledge needed to solve the problem might not contain a representative sample of the kinds of knowledge needed to succeed on other problems of the same type. Further, Messick noted that complex tasks contain multiple constructs that might create performance artifacts that are confounded with the domain of interest (what he called construct-irrelevant variance). For example, a written assessment of problem solving might have construct irrelevant variability related to writing ability. Further, there may be elements of that problem that are

irrelevant to the goals of instruction. For instance, a child may not have prerequisite prior knowledge for solving the problem that is not relevant to the problem-solving construct of interest and may therefore perform poorly because it. Messick was also concerned that if students learn only to master a particular task, their performance might not generalize to other classes of tasks. Using a very obvious example, if we were to teach children the specific vocabulary on a standardized vocabulary measure, we do not get a valid measure of a child's general vocabulary from the test. We no longer have a representative sampling of the child's vocabulary, but a specific measure of words he or she was taught. While under ideal circumstances, Frederiksen and Collins' approach can handle issues related to task specificity, Messick argued that their model does not detail how to deal with construct under representation or problems with task specificity. Further, he pointed out that while under ideal circumstances learning may be a consequence of a systematically valid assessment, there are plenty of elements in the educational system that might make learning from assessment more difficult. For instance, the best intentions for assessment for learning may fade when the focus is on preparation for high-stakes tests. High-stakes testing provides another instance of assessment consequence that potentially competes with the learning consequences. Learning the specific question types or teaching testwiseness may not lead to the generalizable learning of the concepts of interest.

Thus Messick viewed learning as one type of consequence of assessment. Learning from assessment is a consequence of being able to glean information from assessments that can improve future learning. He emphasized the need to consider construct validity, the representativeness of the tasks used, and the context in which the assessment takes place. He also expressed concerns over notions like "authentic" assessment that involve validity claims about certain kinds of tasks being more supportive of learning consequences without providing the means or evidence to support their validity as measures of particular constructs.

A Revised Validity Model on Assessment for Learning

Both Frederiksen and Collins and Messick provide the beginnings of models for describing validity issues in assessment for learning. Frederiksen and Collin's notion that systematically valid assessments that are comprehensible by teachers and students facilitate learning in content domains provides a prototype for assessment for learning situations. Messick's concept of consequential validity provides a lens for viewing learning consequences in light of other consequences associated with assessment. Further, he points out the importance of considering construct validity in the choice of tasks. Below, I build on these ideas to consider a framework for evaluating the validity of assessment for learning contexts.

In describing the framework, I want to note that I take a broad view of what assessment for learning can be. Consistent with my earlier discussion on the nature of assessment for learning, I view assessment for learning as a potential outcome in a variety of situations. The amount of learning from assessment is determined by a complex of variables that can facilitate or hinder teacher and ultimately student learning from assessments. Second, it is important to note that the components in the framework I have developed are overlapping. They provide different perspectives to view the situation from, but are interrelated with one another. Finally, I assume that the situations where assessment for learning is relevant to evaluate are ones where assessment for learning is an intention of the instructional program. One might see the potential for assessment for learning as an unintended consequence, but when evaluating programs to determine

whether assessment for learning is taking place, the focus for the framework is on situations where assessment for learning is an intended outcome. Below, the components of this framework are described.

Component 1: Meaning of the Assessment to the Teacher and Learner

One of the most important elements to creating a valid assessment for learning context is the need to consider how learners and teachers view the assessment context. Because most formal assessments have summative consequences for learners, it is sometimes difficult for learners and teachers to focus on learning rather than on the score from the assessment. Criticism of the model adopted in the United Kingdom has come in this form (e.g., see Ward, 2008). The artifacts from an assessment (e.g., score sheets, rubrics, and so on) can also influence the meaning of assessments regardless of the stated purpose of the assessment. For example, a few years ago I was working with a school district that was using a Rasch-based test that yielded scores that could be used formatively. These Rasch scores had to be paired, however, with a set of skills that could be understood only in the context of the use of additional materials that could be purchased from the test developer. The district did not purchase these and consequently was at a loss as to what to do with the scores of a midyear test. Numeric scores have propensity to focus students on grades. The focus on grades tends to interfere with their ability to focus on learning.

Component 2: The underlying Learning Model of the Assessment

Cognitive science (e.g., Mislevy, 2006; Pellegrino et al., 2001) has done a great deal to build ways of assessing the mental models of tasks that students possess. Analysis of expertise (e.g., Chi, 2006; Ericsson & Ward, 2007), computer simulations of tasks, and componential analyses of the cognitive processes associated with task performance all provide models of learning within domains that have the potential to give teachers and learners information about where students are and how to move to the next level. Hence, a second component in evaluating whether an assessment can support learning is whether there is a model of learning in the domain associated with the assessment that can be used to understand how to improve performance.

Another important issue in dealing with the model of learning is that Messick's (1994) concern over the construct validity of assessments can be assuaged somewhat by considering the model of learning and expertise in the field as the basis for creating assessment tasks. The model of how students acquire expertise in the domain provides the basis for the class of tasks that need to be mastered (Frederiksen & Collins, 1989, 1996). This is an important part of Frederiksen and Collins's model that is crucial in designing direct and authentic assessments.

Component 3: The Assessment Task

This component is one that seems obvious, but does merit some separate discussion. As Messick (1994) noted, the issue of what constitutes an authentic task or one that can facilitate learning is a claim subject to validation. Frederiksen and Collins (1996) suggested that tasks be authentic to the domain (something that is a real task carried out by people in the field) and that it be direct (an actual task rather than something like a multiple choice test that may not apply the knowledge in a meaningful way). While most proponents of assessment for learning argue for task authenticity, it is conceivable that multiple choice or more traditional test like tasks could still provide data that can facilitate learning

Messick's focus on the representativeness and construct validity elements do provide an important basis for assessing whether a task will

facilitate more than specific task performance. Further, Messick's concern about the multiple constructs being assessed when considering complex tasks needs to be addressed. However, that multiple skills are addressed by a task should not disqualify tasks. For one thing, expertise in a domain often requires multiple types of expertise. For example, scientists need not only content knowledge to write a report, but also skill in communication and writing. Further, if there is task-irrelevant variance that leads to potentially incorrect interpretations, one could provide documentation or alternative ways of exploring task performance to work around potentially incorrect interpretations. For instance, if prior knowledge is required to succeed on a task and some students do not have that prior knowledge, one could provide documentation to ensure that teachers provide that knowledge to students who do not have it.

Finally, it is important to note that assessment for learning might be enhanced by exposure to more than one task. Hence, multiple tasks rather than a single task could provide more opportunity for student learning and transfer. For example, Bransford and Schwartz (1999) noted that when viewed as preparation for future learning, a great deal of transfer can be facilitated by providing students with contrasting cases. Including multiple tasks also assuages the concern that Messick had over single-performance assessments providing limited information about learning in a domain. Multiple cases can provide students and teachers with information that goes beyond single task assessments about how students know the domain rather than the task.

Component 4: The Participants Involved in Using the Information from the Assessment

As noted earlier, assessment for learning can occur either directly from a student gleaning information from assessment artifacts, the teacher learning about the student which in turn leads to student success, from interactions among students around an assessment artifact, or from some combination of all of the above. Further, it is possible that learning can occur from interactions of a student with intelligent tutoring systems that adjust instruction on the basis of ongoing assessment information gleaned from students' responses to the tasks presented in the system. It is necessary to examine all of the other components in the system of assessment to determine if the parties involved in the assessment for learning situation can potentially benefit in the ways that are intended. Thus, assessment for learning situations need to consider the individuals involved, their developmental level, and their educational history to determine the meaning of an assessment for an individual (Mislevy, 2007).

Component 5: The Appropriateness of the Information in the Assessment Artifact(s)

Information needs to be present for the assessment to facilitate learning. For example, a set of numeric scores without any other information pertaining to the concepts and procedures students were to learn will not provide the appropriate information for learning from the assessment. Explanations need to be at the appropriate level of expertise for the user (the student and/or teacher). This leads to the next component referring to the comprehensibility of assessment information.

Component 6: Comprehensibility of the Assessment Artifacts

The comprehensibility of the assessment information portrayed in assessment artifacts or verbal descriptions is an important element of the degree that learning can take place. For example, a teacher who does not have the conceptual knowledge to interpret an assessment artifact may be unable to determine how that assessment information can help a student learn. Students who do not know how to interpret a rubric or set of scores

appropriately will not be able to find anything useful for improving performance or learning.

There are many elements that influence how individuals comprehend information from assessments (Shute, 2008). Who needs to comprehend depends upon the nature of the interaction context. For example, students who will be using assessment information to self-assess need information geared at their developmental levels. Students' interactions with teachers who are attempting to facilitate assessment for learning require that the information provided be within the student's zone of proximal development (Vygotsky, 1978) to benefit from the information. Further, the cognitive load (Sweller, 2006) associated with assessment information can impact the interpretability of the data by students and teachers.

Component 7: The Ability to Translate Assessment Information into Actions that Facilitate Learning

Obviously, if the assessment information is not comprehended sufficiently by at least one of the parties involved in an assessment situation, no plan for facilitating learning will be successful. If the teacher understands the outcome of an assessment, then the student can potentially benefit from the teacher's action plan. The process might breakdown if the teacher does not have sufficient knowledge about learning in a domain to create an action plan. Students who are self-assessing need the metacognitive knowledge to figure out how to improve their own learning. Computer tutorials provide decision rules that translate information into instructional actions by the system. The quality of the outcomes depends on how well the decisions match with a valid model of learning for a domain.

Component 8: The Interactions among Participants that Facilitate Learning

Most situations where assessment for learning may take place involve interactions between people. There may be some situations where students completely self-assess, but for the most part, assessment for learning contexts involve interactions between teachers and students. The success of those interactions reflects a variety of factors. For example, power relations between individuals and their past histories together can have an impact on how assessment information is transacted between parties. For instance, a student with a past history of failure in conjunction with a teacher might focus on ability information or interpersonal information rather than what can be gained from an assessment situation. Students working together might be more likely to value assessment information from students they perceive of as knowledgeable and discount information from students who do not. Further, students' interpretation of negative assessment feedback from teachers or from other students can prove problematic. If feedback is provided in a culture that values performance improvement, it is more likely to be successful than if the focus is a grades or other extrinsic rewards (Marshall, 1988; Shute, 2008).

Component 9: The Technical Quality of the Assessments

The last element of the framework concerns issues related to the technical quality of assessments in the context of assessment for learning. As noted by Brookhart (2003), Moss (2003, Shepard (2006), and Smith (2003), aspects of traditional measurement theory concerning technical quality are difficult to apply in assessment for learning situations. As noted earlier, the task and learning theory that are associated with the task are important to validate. Hence, a research base for the model of the task and the interventions associated with

learning from assessments requires a research base to support its use would provide validity evidence. Agreement of experts in the domain concerning the validity of the task also provides evidence. Traditional correlations with other measures also can play a role. Smith notes that a criterion for determining the reliability of measures that are used for formative assessment is the sufficiency of the information provided by the assessment. The validity of performance rubrics are can be determined by expert agreement about the appropriate cutoff for different levels of performance. This can be done through traditional methods such as the Angoff method (Schultz & Whitney, 2004) or can be facilitated through item-response theory models. Reliability of the scoring of performance tasks can be determined by studies of interrater agreement.

One other element that needs to be studied is the reliability and validity of assessment for learning conclusions drawn by teachers and/or students from assessment information. If the evidence examined does not lead learners in the appropriate direction, then the facilitative effect of assessment for learning might be lost. The process of reliably and validly drawing conclusions especially has to be studied in light of different kinds of learners and different learning contexts. Just as traditional tests may show differential item bias, information provided in assessment for learning situations may have differential effectiveness for different learners.

Needs for Further Development in the Framework

The framework developed in this manuscript is a beginning for dealing with the complexity of assessment for learning situations. There are limitations in its applicability that still need to be fleshed out. First, it may be difficult for this framework to describe the spontaneous assessment for learning situations that arise per the coincidence of the right kind of question paired with an answer provided by a student that exposes that student's need to learn more to a teacher. That teachable moment is difficult to validate, but important to student learning. Further, while I have described a number of components that are viable concerns for assessment for learning situations, the framework needs further shaping. In particular finding ways to measure the effect of each of the components is important to using them to evaluate assessment for learning situations. Additionally, more work needs to be done to specify these components in actual rather than hypothetical ways. Lastly, I hope to integrate the work presented here more completely with the work of others in the field (e.g., Shepard, 2006) who have spent their careers determining better ways for teachers to use assessment to improve student learning.

References

Black, P., & Wiliam, D. (1998). Inside the black box. *Phi Delta Kappan*, *80*, 139.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*, 61-100.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement Issues and Practice*, *22*(4), 5-12.

Chi, M. (2006). Laboratory methods for assessing experts' and novices' knowledge. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 167-184). New York: Cambridge University Press.

Ericsson, K., & Ward, P. (2007). Capturing the naturally occurring superior performance of experts in the laboratory: Toward a science of expert and exceptional performance. *Current Directions in Psychological Science*, *16*, 346-350.

Frederiksen, J., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18(9)*, 27-32.

Frederiksen, J. R., & Collins, A. (1996). Designing an assessment system for the workplace of the future. (pp. 193-221). In L. B. Resnick, J. Wirt, & D. Jenkins (Eds.), *Linking school and work: Roles for standards and assessment*. San Francisco: Jossey-Bass.

Mansell, W. (2008, June 20). Every school to get a champion of assessment for learning. *Times Education Supplement*, p. 7.

Marshall, H. (1988). In pursuit of learning-oriented classrooms. *Teaching and Teacher Education*, *4*, 85-98.

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

Messick, S. (1996). *Validity and washback in language testing*. Research Report, Educational Testing Service. (ERIC Document Reproduction Service No. ED403277).

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. I. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 257-305). Westport, CT: American Council on Educational Measuerment/Praeger.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, *36*, 463-469.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational measurement: Issues and Practice, 22(4)*, 13-25.

Pellegrino, J., Chudowsky, N, & Glaser, R. (2001*). Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Schultz, K. S., & Whitney, D. J. (2004). *Measurement theory in action*. Thousand Oaks, CA: Sage.

Shepard, L. (2006). Classroom assessment. In R. I. Brennan (Ed.), *Educational measurement* (4th ed.)(pp. 623-646). Westport, CT: American Council on Educational Measuerment/ Praeger.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practices, 22(4)*, 26-33.

Shute, V. J. (2008). Focus on formative assessment. *Review of Educational Research*, *78*, 153-189.

Sweller, J. (2006). Discussion of "Emerging topics in cognitive load research: Using learner and information characteristics in the design of powerful learning environments." *Applied Cognitive Psychology*, *20*, 353-357.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard.

Ward, H. (2008, October 17). Assessment for learning has fallen prey to gimmicks, says critic. *Times Education Supplement*, p. 18.