KEYWORDS

Simply Assessment

Valerie J. Shute
Educational Psychology and Learning Systems
Florida State University
vshute@fsu.edu

IJLM

Introduction

> *Standardized tests are monstrously unfair to many kids. We're creating a one-size-fits-all system that needlessly brands many young people as failures, when they might thrive if offered a different education whose progress was measured differently.*
>
> *—Robert Reich*

Assessment gets a bum rap. Part of this is because people tend to equate *assessment* with *testing*.[1] Assessment has historically acted as a barrier rather than a bridge to educational opportunity. Suppose you surveyed a random sample of people on the street regarding their feelings about "assessment." Many of them may view it negatively—as unfair, difficult, confusing, inauthentic, boring, constraining, contrived, old school, and so on. Similarly, if you surveyed a random sample of K-12 teachers, many of them, too, may harbor some ill will toward the topic of assessment. Their misgivings may be colored by the No Child Left Behind (NCLB 2002) initiative, which has thus far failed to live up to its promises. NCLB, with its focus on accountability, has promoted "teaching to the test," where ultimately what gets left behind is deeper, more meaningful learning (i.e., knowledge, skills, concepts, and beliefs that are fully understood and can be related to other concepts). Meaningful learning is a desirable goal but much harder to test than rote learning, which is less desirable but fairly easy to test.

There is, however, a more attractive face of assessment, where the primary goal is to improve people's learning (Black and Wiliam 1998; Shute 2007; Stiggins 2008). Stiggins (2008) suggests that we assess for two reasons: to gather evidence to inform instructional decisions, and to encourage learners to try to learn. It is this face of educational assessment that I find to be

Visit IJLM.net

doi: 10.1162/ijlm.2009.0014

Volume 1, Number 2

exciting, powerful, and absolutely critical to support the kinds of learning outcomes and processes necessary to succeed in the 21st century. I'm referring to "formative assessment," which may be thought of as assessment *for* learning, in contrast to "summative assessment" (or assessment *of* learning).

The primary premise underlying this essay is that assessment results can and should have important implications for instruction, positively influencing both the teaching and learning sides of the equation. In today's classrooms, however, assessment is too often used for purposes of grading, promotion, and placement, but not for learning. The stance I take on assessment is that it should: (a) support, not undermine, the learning process for learners and teachers/mentors; (b) provide more formative, compared to summative, information (i.e., give useful feedback during the learning process rather than a single judgment at the end); and (c) be responsive to what is known about how people learn, generally and developmentally.

This essay consists of three parts. First, I broadly define assessment-related terms in relation to the critical roles they have traditionally played in education. Next, I distinguish between the different uses of assessment (i.e., summative and formative). Finally, I describe an evidence-based approach to assessment design that can be used for developing excellent assessments for summative or formative purposes. This approach allows one to measure what learners know, what they believe, and what they're able to do, and facilitates accurate diagnoses to guide learning. I conclude with ideas for incorporating evidence-based assessments into multimedia systems (e.g., games, simulations, web-based learning environments) to support learning of all types and within multifarious contexts, from formal school environments to more informal, out-of-school activities.

Definitions

The goal of this section is to briefly define and disambiguate important assessment terms that often get confounded. For instance, what exactly is the difference between "measurement" and "assessment"? Let's start with the basic idea of measurement. Whenever you need to measure something accurately, you probably grab an appropriate tool to determine how heavy, light, tall, short, fast, slow, hot, cold, bright, dark, straight, or curved something is. We measure to obtain information (data) that may or may not be useful, depending on the accuracy of the tools we use as well as our skill at using them. A measurement such as a person's height, a room's temperature, or a car's speed is not an assessment but a piece of data in a standardized unit. How does this relate to education?

Educational Measurement

Measurements are not to provide numbers but insight.
—Ingrid Bucher

Educational measurement, in the context of this essay, refers to the application of a measuring tool (or standard scale) to determine the degree to which educationally valuable knowledge, skills, and other attributes have been or are being acquired. It thus entails the collection and analysis of learner data. According to the National Council on Measurement in Education website, this includes theory, techniques, and instrumentation available for measurement of educationally relevant human, institutional, and social characteristics. A *test* is education's equivalent of a ruler, thermometer, or radar gun. But note that a test does not improve learning any more than a thermometer cures a fever; both are simply tools. Moreover, as Snow and Jones (2001) point out, tests alone cannot enhance educational outcomes. Rather, tests—assuming they are valid and reliable—can guide improvement if they motivate adjustments to the educational system (i.e., provide the basis for bolstering curricula, ensure support for struggling learners, guide professional development opportunities, and distribute limited resources fairly).

Again, we measure things in order to get information, which may be quantitative or qualitative.[2] How we choose to *use* the data is a different story. For instance, back in the early 1900s, students' abilities and intelligence were extensively measured. However, this wasn't done to help them learn better or to progress. The main purpose of testing was to track students into appropriate paths, with the understanding that their aptitudes were inherently fixed. A dominant belief during that period was that intelligence was part of a person's genetic makeup, thus testing was aimed specifically at efficiently assigning students to high, middle, or low educational tracks according to their supposedly

innate mental abilities (Terman 1916). There was a fundamental shift to practical education in the country during the early 1900s, countering "wasted time" in schools and abandoning the classics as useless and inefficient for the masses (Shute 2007). Early educational researchers and administrators inserted into the national educational discourse the metaphor of the school as a "factory" (Kliebard 1987), and that metaphor has persisted to this day.

Assessment

> *Assessment should not merely be done to students; rather, it should also be done for students, to guide and enhance their learning.*
>
> *—NCTM (2000)*

Assessment involves much more than just measurement. That is, in addition to systematically collecting and analyzing information, it also involves interpreting and acting on information about learners' understanding and/or performance in relation to educational goals. Measurement, then, can be viewed as a subset of assessment.

Assessment information may be used by a variety of stakeholders (teachers, administrators, students, parents, etc.) and for a variety of purposes, such as to improve learning outcomes, programs, and services, and also to establish accountability. Furthermore, there is an assortment of procedures associated with the different purposes. For example, if your goal was to enhance an individual's learning and you wanted to determine her progress toward an educational goal, you could: (a) administer a quiz; (b) view a portfolio of her work; (c) ask the student (or peers) to evaluate her progress; (d) watch the person solve a complex task; (e) review her lab reports or journal entries; and so on.

In addition to having different purposes and procedures for obtaining information, assessments may also be differentially referenced or interpreted, in relation, for example, to normative data or to a criterion. *Norm-referenced* interpretation compares learner data to that of other individuals or to a larger group but can also involve comparisons to oneself (e.g., asking a person how she's feeling and getting a "Better than usual" response is a norm-referenced interpretation). The purpose of norm-referenced interpretation is to establish what is typical or reasonable. On the other hand, *criterion-referenced* interpretation involves establishing what a person can or cannot do, or typically does or does not do—specifically in relation to a criterion. Note that if the purpose of the assessment is to support personal learning, then criterion-referenced interpretation is required (for more information see Nitko 1980).

The final general assessment issue has to do with *who* is doing the assessing. Very often, it is the teacher. However, *self-assessment* may be a viable option, as well as an important skill, especially if a valued educational goal is to produce self-directed and productive lifelong learners. Promoting learners' self-assessment in relation to setting reasonable learning goals involves supporting knowledge of specific goals and learners' progress toward them. It also involves supporting learners' metacognitive skills of reflection and revision. Alternatively, *peer assessment* involves individuals collaborating with one another to solve, explain, or understand a problem or task. There are a variety of benefits (e.g., cognitive, social, motivational) from encouraging learners to work collaboratively. An effective teacher should emphasize a high and equal level of interaction among group members, giving all an opportunity to negotiate meaning, acquire new strategies and skills, and develop higher-order thinking skills. However, as collaboration becomes an increasingly important aspect of 21st-century learning, this introduces not only opportunities but also serious challenges to assessment that will need to be resolved with innovative research (Jeong 2005; Macdonald 2003; Shute, Jeong, and Zapata-Rivera, in press; Shute, Jeong, Spector, Seel, and Johnson, in press).

Measurements of Assessment Quality

Because assessment is a process by which information is obtained relative to a known objective, and since inferences are made about what a person knows (unobservable) on the basis of responses to assessment tasks (observable), there's always some uncertainty in inferences made on the basis of assessments. So an important goal in educational measurement is to collect really good information about the learner(s) and to minimize uncertainty or error. Consequently, key aspects of assessment quality are consistency and validity.

Consistency

The broad term *consistency* is used in this essay rather than the more familiar term *reliability* because it includes not only the quantitative aspects of reliability

(e.g., correlations between parallel forms of tests) but also qualitative aspects of assessment (e.g., consistency in a teacher's description of a learner's performance on two comparable tasks). To illustrate, consider the produce scale at your local grocery store. If you weigh two pounds of broccoli and the scale is reliable, the same scale should register the same weight for the same broccoli an hour later. Similarly, classroom tests and standardized exams should be stable, and it shouldn't make much difference whether a learner takes the test at 10:00 AM or at 2:00 PM. Another measure of consistency—internal consistency—relates to the items within a test. For instance, if you create an Algebra 1 test, you'd assume that if a learner correctly solves a difficult linear equation problem, then he or she will solve easier linear equation problems correctly. Similarly, the notion of generalizability is often used with performance assessments and portfolios, which addresses the adequacy with which you can generalize from a randomly sampled set of observations to a universe of observations.

Validity

There are a number of different types of validity. In general, *validity* refers to the extent to which the assessment accurately measures what it is supposed to measure, and the accuracy of the inferences made from test results. For instance, if you wanted to assess learners' math problem-solving skills, but you gave them a personality inventory to complete, that would not result in a valid assessment (and you should probably find a new day job). Even if an assessment is judged to be consistent and stable (see foregoing paragraph), it may not, in fact, be a valid measure. Let's use a scale analogy again, only now it's your bathroom scale. Suppose you step on your scale 10 times in a row and your scale indicates, without fail, that you weigh 150 pounds. The *consistency* of your scale may be very good, but it may not be accurate (valid) if you actually weigh 165 pounds. Because teachers, parents, school districts, and so on currently make decisions about learners based on assessment results (e.g., grades, retention, graduation), the validity inferred from the assessments is essential, and it's even more crucial than the consistency. So, consistency is a prerequisite for validity. That is, inconsistency in observations always threatens their validity. On the other hand, simply having consistency in what is observed does not ensure the validity of those observations.

Uses of Assessment

There are various types of assessment, often presented in contrast to one another (e.g., summative vs. formative assessment). The choice and use of a particular type of assessment depends on the educational purpose. Schools generally make heavy use of summative assessment (also known as assessment *of* learning), which is useful for accountability purposes (e.g., unidimensional assessment for grading and promotion purposes) but only marginally, if at all, useful for supporting personal learning. In contrast, learner-centered measurement models rely mostly on formative assessment, also known as assessment *for* learning, which can be very useful in guiding instruction and supporting individual learning but may not be particularly consistent or valid. One current downside of the assessment-for-learning model is that it is often implemented in a nonstandardized and hence less rigorous manner than summative assessment, and thus can hamper the validity and consistency of the assessment tools and data (Shute and Zapata-Rivera, in press). This is not to say such assessments don't have value. Rather, it's a call for research to come up with new measures or techniques to determine assessments' value and/or utility (e.g., a meta-analysis approach using many formative assessments to provide an aggregate picture that cannot be seen clearly through individual assessments). Strong formative assessment research is urgently needed given changes in (a) the types of learning we are valuing today (and in the near future), as well as (b) the new, broader set of contexts in which learning is taking place.

Summative and Formative Assessment

> *When the cook tastes the soup, that's formative; when the guests taste the soup, that's summative.*
>
> *—Robert Stake*

The two most familiar types of assessment are summative and formative. *Summative assessment* reflects the so-called traditional approach used to assess educational outcomes. This involves using assessment information for high-stakes, cumulative purposes, such as for grades, promotion, certification, and so on. It is usually administered after some major event, like the end of the school year or marking period, or before a big event, like college entry. Benefits of this approach include the following: (a) it allows for comparing learner performances across diverse populations on

clearly defined educational objectives and standards; (b) it provides reliable data (e.g., scores) that can be used for accountability purposes at various levels (e.g., classroom, school, district, state, and national) and for various stakeholders (e.g., learners, teachers, and administrators); and (c) it can inform educational policy (e.g., curriculum or funding decisions). *Formative assessment* reflects a more progressive approach in education. This involves using assessments to support teaching and learning. Formative assessment is incorporated directly into the classroom curriculum and uses results from learners' activities as the basis on which to adjust instruction to promote learning in a timely manner. A simple example would be a teacher giving a "pop quiz" to his students on some current event, immediately analyzing their scores, and then refocusing his lesson to straighten out a prevalent misconception shared by the majority of students in the class. This type of assessment is administered more frequently than summative assessment, and has shown great potential for harnessing the power of assessments to support learning in different content areas and for diverse audiences (Black and Wiliam 1998; Hindo, Rose, and Gomez 2004; Schwartz, Bransford, and Sears 2005). In addition to providing teachers with evidence about how their classes are learning so that they can revise instruction appropriately, formative assessment directly involves learners in the process, such as by providing feedback that will help them gain insight into how to improve, and by suggesting (or implementing) instructional adjustments based on assessment results.

Table 1 characterizes four assessment variables (main role in the classroom, frequency of administration, typical format, and feedback) that are characteristic of summative and formative assessment. Note that neither type of assessment is an educational panacea—both have enormous strengths and serious limitations. Elsewhere (Shute 2007) I have suggested merging the best features from each into a unified and more powerful educational approach. Table 1 is intended to convey general aspects of each approach in terms of the variables and should not be viewed as establishing definitive categorizations.

The preceding definitions are intended to lay the foundation for understanding the "what" part of assessment. To accomplish the important goal of developing really good assessments that can also support learning, I now present the "how" part of the story; namely, an overview of evidence-centered design

Table 1 Assessment Variables in Relation to Summative and Formative Approaches

Variables	Summative Assessment	Formative Assessment
Role of assessment	Assessment *of* learning, to quantify fixed and measurable aspects of learners' knowledge, skills, and abilities. Used for accountability purposes, often with norm-referenced tests. Produces a static/snapshot of the learner.	Assessment *for* learning, to characterize important aspects of the learner. The main focus is on aspects of learner growth, employing criterion-referenced tests, used to help learners learn and teachers teach better.
Frequency of assessment	Infrequent, summative assessments using standardized tests. The focus is on product or outcome (achievement) assessment. Such tests are typically conducted at the end of a major event (e.g., unit, marking period, school year).	Intermittent, formative assessment. The focus is more process oriented (but needn't exclude outcomes). Assessments of this type are administered as often as desired and feasible—monthly, weekly, or even daily. Administration is informal.
Format of assessment	Objective assessments, often using selected responses. The focus is on whether the test is valid and consistent more than the degree to which it supports learning.	Constructed responses and an authentic context, collected from multiple sources (e.g., quizzes, portfolios, self-appraisals, and presentations).
Feedback	Correct or incorrect responses to test items and quizzes, or just overall score. Support of learning is not the intention.	Global and specific diagnoses, with suggestions for ways to improve learning and teaching. Feedback is helpful, rather than judgmental.

Note. This table is adapted from Shute (2007).

(ECD), which supports the design of valid assessments. ECD entails developing competency models, scoring rules, and associated assessments.

Evidence-Centered Design

> *The nature of the construct being assessed should guide the selection or construction of relevant tasks, as well as the rational development of construct-based scoring criteria and rubrics.*
>
> *—Sam Messick*

The fundamental ideas underlying ECD came from Messick (1994; see quote above). This process begins by identifying what should be assessed in terms of knowledge, skills, or other attributes. These variables cannot be observed directly, so behaviors and performances that demonstrate these variables should be identified instead. This is followed by determining the types of tasks or situations that would draw out such behaviors or performances. An overview of the ECD approach is described below (for more on the topic, see Mislevy and Haertel 2006; Mislevy, Almond, and Lukas 2004; Mislevy, Steinberg, and Almond 2003).

ECD Models

Again, the primary purpose of an assessment is to collect information that will enable the assessor to make inferences about learners' competency states—what they know, believe, can do, and to what degree. Accurate inferences of competency states support instructional decisions that can promote learning. ECD defines a framework that consists of three theoretical models that work in concert. The ECD framework allows/requires an assessor to: (a) define the claims to be made about learners' competencies, (b) establish what constitutes valid evidence of the claim, and (c) determine the nature and form of tasks that will elicit that evidence. These three actions map directly onto the three main models of ECD shown in figure 1.

A good assessment has to elicit behavior that bears evidence about key competencies, and it must also provide principled interpretations of that evidence in terms that suit the purpose of the assessment. Working out these variables, models, and their interrelationships is a way to answer a series of questions posed by Messick (1994) that get at the very heart of assessment design.

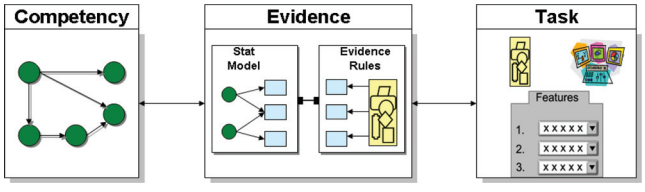

Figure 1 Three main models of an evidence-centered assessment design.

Competency Model

What collection of knowledge, skills, and other attributes should be assessed? This can be rephrased as: What do you want to say about the person at the end of the assessment? Variables in the competency model (CM) are usually called "nodes" and describe the set of person variables on which inferences are based. The term *student* (or *learner*) *model* is used to denote an instantiated version of the CM, like a profile or report card, only at a more refined grain size. Values in the learner model express the assessor's current belief about the level of each variable within the learner's CM. For example, suppose the CM for a science class that valued the general competency of systems thinking contained a node for "Create a causal loop diagram." The value of that node—for a student who was really facile at understanding and drawing causal loop diagrams—may be "high" (if the competency levels were divided into low, medium, and high), based on evidence accumulated across multiple, relevant tasks.

Evidence Model

What behaviors or performances should reveal differential competency levels? An evidence model (EM) expresses how the student's interactions with, and responses to, a given problem constitute evidence about competency model variables. The EM attempts to answer two questions: (a) What behaviors or performances reveal targeted competencies; and (b) What's the connection between those behaviors and the CM variable(s)? An EM lays out the argument about why and how observations in a given task situation (i.e., student performance data) constitute evidence about CM variables. Using the same node as illustrated in the CM section above, the EM would clearly indicate the aspects of causal loop diagrams that must be present (or absent) to indicate varying degrees of understanding or mastery of that competency. The same logic/methods apply to noncognitive variables as well, stating clearly the rubrics for scoring aspects of creativity, teamwork, etc.

Task Model

What tasks should elicit those behaviors that comprise the evidence? A task model (TM) provides a framework for characterizing and constructing situations with which a learner will interact to provide evidence about targeted aspects of knowledge or skill related to competencies. These situations are described in terms of: (a) the presentation format (e.g., directions, stimuli), (b) the specific work or response products (e.g., answers, work samples), and (c) other variables used to describe key features of tasks (e.g., knowledge type, difficulty level). Thus, task specifications establish what the learner will be asked to do, what kinds of responses are permitted, what types of formats are available, and other considerations, such as whether the learner will be timed, allowed to use tools (e.g., calculators, dictionaries), and so forth. Multiple task models can be employed in a given assessment. Tasks are the most obvious part of an assessment, and their main purpose is to elicit evidence (which is observable) about competencies (which are unobservable).

Design and Diagnosis

As shown in figure 1, assessment design flows from left to right, although in practice it's more iterative. Diagnosis (or inference) flows in the opposite direction. That is, an assessment is administered, and the learners' responses made during the solution process provide the evidence that is analyzed by the evidence model. The results of this analysis are data (e.g., scores) that are passed on to the competency model, which in turn updates the claims about relevant competencies. In short, the ECD approach provides a framework for developing assessment tasks that are explicitly linked to claims about personal competencies via an evidentiary chain (e.g., valid arguments that serve to connect task performance to competency estimates), and are thus valid for their intended purposes.

The following section describes some ideas that involve embedding ECD-based assessments within multimedia environments, such as games and simulations.

Stealth Assessment

When assessment is so seamlessly woven into the fabric of the learning environment that it is virtually invisible, I call this *stealth assessment*. This kind of assessment is intended to support learning and remove (or seriously reduce) test anxiety, while not sacrificing validity and consistency (see Shute, Hansen, and Almond 2008). Stealth assessment is accomplished via automated scoring and machine-based reasoning techniques to infer things that would be too hard for humans to infer, such as concurrently estimating values of multiple competencies across a network of skills for numerous learners. One good technique for accomplishing these inferencing goals involves using what are called Bayesian networks (or "Bayes nets," for short).

In learning environments (e.g., online games) with stealth assessment, and in line with the discussion on ECD above, the CM accumulates and represents belief about targeted aspects of knowledge or skill, expressed as probability distributions for CM variables (Almond and Mislevy 1999). Evidence models identify what the learner says or does that can provide evidence about those skills (Steinberg and Gitomer 1996) and express in a psychometric model how the evidence depends on the CM variables (Mislevy 1994). Action models express situations that can evoke required evidence. Key elements of the approach include (a) employing ECD, which systematically analyzes the assessment argument, including the claims to be made about the learner and the evidence that supports (or fails to support) those claims, and (b) formative assessment to guide instructional experiences.

To illustrate the stealth assessment idea, and as part of ongoing research in this area, my colleagues and I recently modeled some 21st-century competencies within game environments, including (a) creative problem-solving (within *The Elder Scrolls IV: Oblivion*, 2006, by Bethesda Softworks; see Shute, Ventura, Bauer, and Zapata-Rivera, in press) and (b) systems thinking. Regarding the latter competency, we recently provided an analysis (or worked example) of an existing 3D immersive game called *Quest Atlantis: Taiga Park* (Barab 2006; Barab et al. 2007). We demonstrated how evidence of the systems-thinking competency may be automatically gathered and interpreted from a learner during the course of game play (for more details, see Shute, Masduki et al., in press). Briefly, we began by developing ECD models relevant to systems-thinking skill. For the worked example, we focused on just one branch of the CM: Model the system. The quests that players undertake in *Taiga Park* occur within five different "Missions," all of which are designed to make learners think carefully about

complex ecological systems—their interconnections and dynamic relations among elements. Thus, the fit between our selected competency and that goal of the game was ideal.

As part of the worked example, we modeled a hypothetical learner (Clara) in terms of her systems-thinking skill at two points in time (i.e., Time 1: an initial quest; and Time 2: a final quest). The example showed quantitative and qualitative changes to her systems-thinking abilities over time. For instance, we compared Clara's causal loop diagrams created at Time 1 and Time 2 to an expert's diagram (note: "Create a causal loop diagram" is one of 24 nodes being estimated in our systems-thinking CM). These comparisons are made possible by automatically standardizing her diagram, and then overlaying the standardized map onto an expert map. The tool that we used for the standardization and comparison is an Excel-based software application called jMap (Jeong 2008; Shute, Jeong, and Zapata-Rivera, in press), designed to accomplish the following goals: (1) elicit, record, and automatically code mental models; (2) visually and quantitatively assess changes in mental models over time; and (3) determine the degree to which the changes converge toward an expert's (for more information about the program as well as relevant papers and links, see: http://garnet.fsu.edu/~ajeong).

Our systems-thinking CM was created after an extensive literature review on the topic. Nodes in the CM were statistically linked to each other in terms of conditional probabilities and comprise different levels in the network. For instance, the "parent" node represents an estimate of the learner's general systems-thinking skill, given all of the evidence collected at that point. Low-level nodes (i.e., those without progeny) are explicitly linked to indicators obtained from the game via our evidence model. Such indicators (e.g., the accuracy of Clara's causal loop diagram in relation to an expert's map, derived from jMap) provide information that "feeds" the Bayes net. Once the information is inserted into the Bayes net, it is propagated throughout the network to all of the nodes, whose estimates are subsequently altered. Figure 2 shows a fragment of the Bayes net, with Clara's estimated competencies at an early stage of learning in *Taiga Park*. That is, Clara's Time 1 estimate for the competency "Create causal loop diagram" is medium; her "elaborate reasoning" competency, however, is estimated at low, as is her overall competency, "model the system." She has more work to do in *Taiga*, and this analysis and diagnosis targets particular areas for improvement.

Finally, information that is obtained from comparing Clara's causal diagram to an expert map clearly demonstrates any misconceptions, and it can be used as the basis for formative feedback that may be presented by the teacher or automatically by the environment. For example, given particular errors of omission apparent in Clara's early map, the system would provide the following feedback: "*Nice job, Clara—but you forgot to include the fact that sediment*

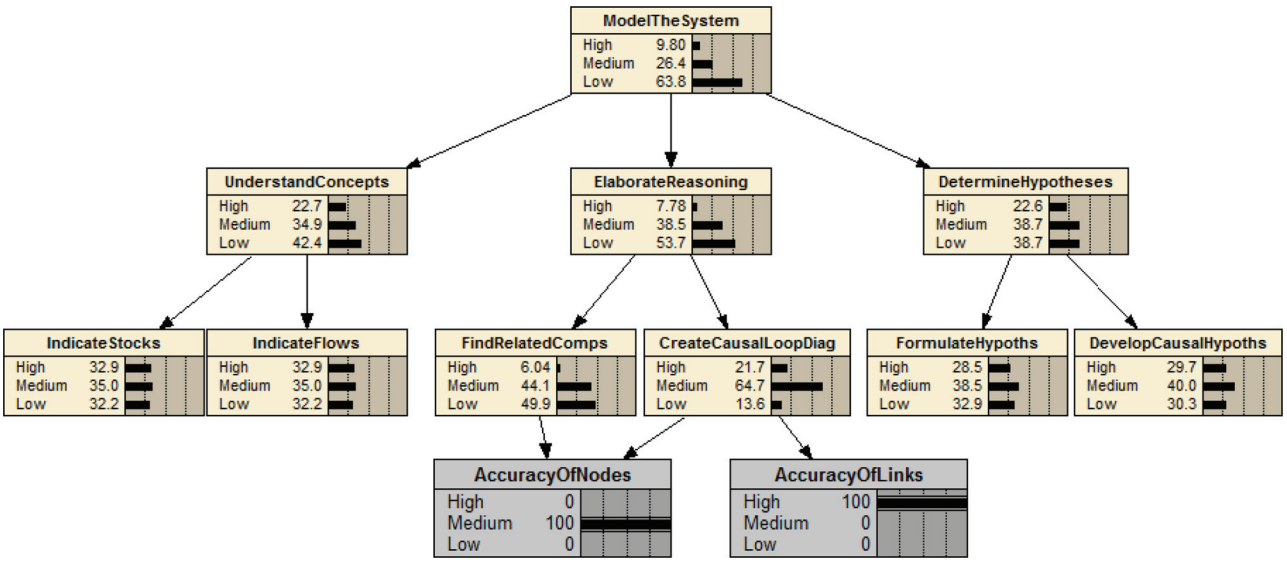

Figure 2 Bayesian model for Clara at Time 1.

increases water temperature, which decreases the amount of dissolved oxygen in the water. That's the reason the fish are dying—they don't have enough oxygen." Moreover, the *Taiga* lab technician (or another knowledgeable character in the park) could provide feedback to Clara in the form of the expert causal loop diagram, explicitly including her omitted variables in the picture. That way, she could see for herself what she'd left out.

Summary and Conclusions

> *All assessment is a perpetual work in progress.*
> —*Linda Suske*

Education, especially in the United States, seriously needs to move into the 21st century. This will entail a re-focus on its primary mission, which is to ensure that individuals learn valuable knowledge and skills—cognitive and noncognitive—to enable them to contribute as well-adapted, effective members of a global society.

Assessment can and should play a critical role in this educational reformulation. Over a dozen years ago, the National Research Council (NRC 1996) made a similar plea, which has yet to be adequately addressed. Table 2 presents a modified version of the NRC call for changes in the focus on assessment needed to support the goal of educational reform for the 21st century.

Regarding the various types of assessment described in this essay, each has a role to play in improving teaching and learning, and needs to be part of a total, balanced assessment system. Using different kinds of assessment will allow us to view learners' knowledge, skills, and other attributes from multiple perspectives, providing a much clearer picture of each learner (Fletcher 2007). The more we know about learners, the better we can provide them with optimal support at the time they really need it. Moreover, it's crucial to involve learners in the assessment process through peer and self-assessment. This stimulates the use of higher-order thinking skills and helps learners to understand why they're learning different things (Shute 2008). Providing frequent and constructive feedback to learners has also been found to significantly improve learning (Kluger and DeNisi 1996; Narciss and Huth 2004; Shute, Hansen, and Almond 2008).

Table 2 Changing Assessment Foci

Less Focus on Assessing	*More* Focus on Assessing
Learning outcomes	Learning processes
What is easily measured	What is most highly valued
Discrete, declarative knowledge	Rich, authentic knowledge and skills
Content knowledge	Understanding and reasoning, within and across content areas
What learners do *not* know	What learners understand and can do
By teachers alone	By learners engaged in ongoing assessment of their work and that of others

Toward a specific set of *principles* of good assessment, I've merged recommendations from Kellough and Kellough (1999), Mislevy, Steinberg, and Almond (2003), and Shute (2008) to yield the following:

- Understand and specify in advance of teaching the achievement targets (i.e., competencies) that learners are supposed to attain.
- Inform the learners, simply and clearly, about the competencies (as well as the associated rubrics), from the very beginning of the teaching and learning process.
- Use classroom assessments to bolster learners' confidence and help them assume responsibility for their own learning, with the goal of engendering lifelong learners.
- Translate assessment results into frequent, descriptive feedback (not judgmental, subjective, or norm referenced), providing learners with specific insights on how to improve.
- Continuously adjust instruction (whether classroom- or computer-based) relative to the results of the formative assessments.
- Engage learners in regular self-assessment with standards held constant so that they can watch themselves grow over time and feel empowered.

In conclusion, I'd argue that the most important and powerful feature of assessment involves using results to make improvements and decisions. This is true whether the assessment is used to support personal learning or for accountability purposes. Another important feature of assessment is to make learning—processes and products—visible to all stakeholders. That is, a person's knowledge (and other mental states and traits) is invisible to others, and sometimes to oneself (e.g., tacit knowledge). ECD-based assessments can contribute to improved teaching and learning,

and can explicate the evidentiary argument supporting various claims.

Knowing *when* to use a particular type of assessment and *how* to interpret the results is not easy. Similarly, designing assessments using an evidence-based approach is non-trivial. But consider the potential end result—assessments that exert substantial influence on the quality of information provided to teachers and learners to support instructional decision making and meaningful learning. This essay has briefly touched on different assessment topics and approaches, calling for a rational understanding of what we value in terms of competencies to be instructed and assessed, both for the present and with an eye toward the future. Knowing what a learner knows comes from obtaining quality evidence, which in turn is obtained from carefully designed assessment tasks. The ideas herein are intended to support teachers and learners, especially when the implemented assessment ideas include sufficient practice opportunities and targeted feedback for learners.

Notes

1. In fact, these are not the same (see Shute and Zapata-Rivera, in press), which I hope to demystify in this essay.
2. For a fuller, more balanced perspective on educational measurement, see Messick (1989) and Oosterhof (2009), which extend educational measurement beyond statistical conceptualizations and numbers to include qualitative information as well.
3. In game environments, I use the term *action model* instead of task model, as defined in the ECD section described earlier. This defines the relevant actions during a quest, as well as each action's associated indicators of success.

References

Almond, R. G., and R. J. Mislevy. 1999. Graphical models and computerized adaptive testing. *Applied Psychological Measurement* 23 (3):223–237.

Barab, S. A. 2006. From Plato's Republic to Quest Atlantis: The role of the philosopher-king. *Technology, Humanities, Education, and Narrative* 2 (Winter):22–53.

Barab, S. A., T. D. Sadler, C. Heiselt, D. Hickey, and S. Zuiker. 2007. Relating narrative, inquiry, and inscriptions: Supporting consequential play. *Journal of Science Education and Technology* 16 (1):59–82.

Black, P., and D. Wiliam. 1998. Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice* 5 (1):7–74.

Fletcher, G. 2007. *Assessing learning from a holistic approach: Creating a balanced system of learning assessment*. Paper presented the Congreso Internacional Evaluacion Factor de Calidad Educativa, Queretaro, Mexico (October 19, 2007).

Hindo, C., K. Rose, and L. M. Gomez. 2004. Searching for Steven Spielberg: Introducing iMovie to the high school English classroom: A closer look at what open-ended technology project designs can do to promote engaged learning. In *Proceedings of the 6th International Conference on Learning Sciences*, 606–609. Mahwah, NJ: Erlbaum.

Jeong, A. 2005. A guide to analyzing message-response sequences and group interaction patterns in computer-mediated communication. *Distance Education* 26 (3):367–383.

Jeong, A. 2008. jMap. http://jmap.wikispaces.com/ (accessed May 5, 2009).

Kellough, R. D., and N. G. Kellough. 1999. *Secondary school teaching: A guide to methods and resources; planning for competence*. Upper Saddle River, NJ: Prentice Hall.

Kliebard, H. 1987. *The struggle for the American curriculum, 1893–1958*. New York: Routledge and Kegan Paul.

Kluger, A. N., and A. DeNisi. 1996. Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin* 119 (2):254–284.

Macdonald, J. 2003. Assessing online collaborative learning: Process and product. *Computers & Education* 40 (4):377–391.

Messick, S. 1989. Validity. In *Educational measurement*, 3rd ed., edited by R. L. Linn, 13–104. New York: Macmillan.

Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher* 32 (2):13–23.

Mislevy, R. J. 1994. Evidence and inference in educational assessment. *Psychometrika* 59:439–483.

Mislevy, R. J., R. G. Almond, and J. Lukas. 2004. *A brief introduction to evidence-centered design*. CSE Technical Report 632. Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

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

Mislevy, R. J., L. S. Steinberg, and R. G. Almond. 2003. On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective* 1 (1):3–62.

Narciss, S., and K. Huth. 2004. How to design informative tutoring feedback for multimedia learning. In *Instructional design for multimedia learning*, edited by H. M. Niegemann, D. Leutner, and R. Brünken, 181–195. Münster, Germany: Waxmann.

National Council of Teachers of Mathematics (NCTM). 2000. *Principles and standards for school mathematics*. Reston, VA: NCTM.

National Research Council (NRC). 1996. *The national science education standards*. Washington, DC: National Academy Press.

Nitko, A. J. 1980. Distinguishing the many varieties of criterion-referenced tests. *Review of Educational Research* 50:461–485.

No Child Left Behind Act (NCLB). 2002. Title 1: Improving the academic achievement of the disadvantaged.

Summary of Final Regulations. http://www.ed.gov/programs/titleiparta/index.html#reg (accessed December 19, 2008).

Oosterhof, A. 2009. *Developing and using classroom assessments*, 4th ed. Upper Saddle River, NJ: Pearson.

Reich, R. 2000. One education does not fit all. *New York Times*, Op-Ed, July 11, 2000.

Schwartz, D. L., J. D. Bransford, and D. L. Sears. 2005. Efficiency and innovation in transfer. In *Transfer of learning from a modern multidisciplinary perspective*, edited by J. Mestre, 1–51. Greenwich, CT: Information Age Publishing.

Shute, V. J. 2007. Tensions, trends, tools, and technologies: Time for an educational sea change. In *The future of assessment: Shaping teaching and learning*, edited by C. A. Dwyer, 139–187. New York: Lawrence Erlbaum Associates, Taylor & Francis Group.

Shute, V. J. 2008. Focus on formative feedback. *Review of Educational Research* 78 (1):153–189.

Shute, V. J., E. G. Hansen, and R. G. Almond. 2008. You can't fatten a hog by weighing it—or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence and Education* 18 (4):289–316.

Shute, V. J., A. C. Jeong, J. M. Spector, N. M. Seel, and T. E. Johnson. In press. Model-based methods for assessment, learning, and instruction: Innovative educational technology at Florida State University. In *2009 Educational Media and Technology Yearbook*, edited by M. Orey. Westport, CT: Greenwood Publishing Group.

Shute, V. J., A. C. Jeong, and D. Zapata-Rivera. In press. Using flexible belief networks to assess mental models. In *Instructional design for complex learning*, edited by B. B. Lockee, L. Yamagata-Lynch, and J. M. Spector. New York: Springer.

Shute, V. J., I. Masduki, O. Donmez, Y. J. Kim, V. P. Dennen, A. C. Jeong, and C.-Y. Wang. In press. Assessing key competencies within game environments. In *Computer-based diagnostics and systematic analysis of knowledge*, edited by D. Ifenthaler, P. Pirnay-Dummer, and N. M. Seel. New York: Springer-Verlag.

Shute, V. J., M. Ventura, M. I. Bauer, and D. Zapata-Rivera. In press. Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In *The social science of serious games: Theories and applications*, edited by U. Ritterfeld, M. J. Cody, and P. Vorderer. Philadelphia, PA: Routledge/LEA.

Shute, V. J., and D. Zapata-Rivera. In press. Educational measurement and intelligent systems. In *Third edition of the international encyclopedia of education*, edited by E. Baker, B. McGaw, and P. Peterson. Oxford, UK: Elsevier.

Snow, C. E., and J. Jones. 2001. Making a silk purse. *Education Week Commentary*, April 25.

Stake, R. (cited in Earl, L., 2004). Assessment as learning: Using classroom achievement to maximize student learning. *Experts in Assessment*. Thousand Oaks, CA: Corwin Press.

Steinberg, L. S., and D. G. Gitomer. 1996. Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science* 24:223–258.

Stiggins, R. 2008. Assessment manifesto: A call for the development of balanced assessment systems. http://www.nmsa.org/portals/0/pdf/advocacy/other_resources/AssessmentManifesto08.pdf (accessed December 21, 2008).

Terman, L. M. 1916. *The measurement of intelligence*. Cambridge, MA: Riverside Press.