Formative and Stealth Assessment

25

Valerie J. Shute and Yoon Jeon Kim

Abstract

Assessing generally refers to the process of gathering information about a person relative to specific competencies and other attributes, in formal or informal learning contexts. This should lead to valid and reliable inferences about competency levels, which in turn may be used for diagnostic and/or predictive purposes. Too often, classroom and other high-stakes assessments are used for purposes of grading, promotion, and placement, but not to enhance learning. In this chapter, we focus on formative assessment which posits that assessment should (a) encourage and support, not undermine, the learning process for learners and teachers; (b) provide formative information whenever possible (i.e., give useful feedback during the learning process instead of a single judgment at the end); and (c) be responsive to what is known about how people learn, generally and developmentally. This type of assessment has as its primary goal improvement of learning, which is critical to support the kinds of learning outcomes and processes necessary for students to succeed in the twenty-first century. It is referred to as "formative assessment," or assessment *for* learning, in contrast to "summative assessment" (or assessment *of* learning). This chapter overviews the role of formative assessment in education generally, and also touches on stealth assessment specifically—an evidence-based approach to weaving assessments directly into learning environments (Shute, Computer games and instruction. Charlotte, NC: Information Age Publishers, 2011).

Keywords

Competency • Evidence-centered design (ECD) • Formative assessment • Stealth assessment

Introduction

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

In the United States, assessment currently plays a significant (and often heavy-handed) role in educational systems. A prevalent form of assessment in education today is the standardized test. The primary goal of standardized tests is to ensure accountability of schools and teachers. Our nation's usage of standardized tests has increased considerably since the No Child Left Behind (NCLB) Act was signed into law in 2001 (Chappius & Chappius, 2008). For example, before NCLB, 19 states required annual reading and mathematics tests in grades 3–8, and one test administered in high school. By 2006, every state required standardized testing (Jennings & Rentner, 2006).

V.J. Shute (✉) • Y.J. Kim
Florida State University, 1114 W. Call Street, PO Box 3064453, Tallahassee, FL 32306-4453, USA
e-mail: vshute@fsu.edu; yjkim.fsu@gmail.com

J.M. Spector et al. (eds.), *Handbook of Research on Educational Communications and Technology*,
DOI 10.1007/978-1-4614-3185-5_25,

Although there is a little evidence supporting positive effects of the NCLB Act, there is extensive criticism about the hidden costs of NCLB. For instance, Stiggins (2002) argued, "*We are a nation obsessed with the belief that the path to school improvement is paved with better, more frequent, and more intense standardized testing. The problem is that such tests, ostensibly developed to 'leave no student behind,' are in fact causing major segments of our student population to be left behind because the tests cause many to give up in hopelessness—just the opposite effect from that which politicians intended.*" (p. 2).

The primary problem with current assessment practices is that the information from the assessment currently is not being used, as it could and should, to support teaching and learning (e.g., Shute, 2007; Symonds, 2004; Wiliam & Thompson, 2007). Typically, classroom assessments are only administered at the end of some major chunk of time with assessment results arriving too late for teachers to effectively act on them. Symonds (2004) highlighted this problem as she explored policies and practices in dozens of schools that were classified into two groups: successful and unsuccessful in closing the achievement gap. The report showed clear, striking differences between the gap-closing versus non-gap-closing groups—particularly with regard to the use of data. Gap-closing schools assessed students often and used the results to make changes in their instructional program. Non-gap-closing schools assessed students infrequently and did not use the data to effect instructional changes. Two recommendations that emerged from the Symonds study (and which have been endorsed by the Council of Chief State School Officers (2004)) are the following: (1) schools need frequent, reliable data, and (2) teachers need support to use data effectively.

Broadly speaking, the type of assessment that uses test information to support learning is called formative assessment. Despite growing evidence that this type of assessment supports student learning, we don't see wide application of formative assessment in classrooms. Two explanations for the limited adoption of formative assessment in the classroom are the following: (a) it's hard to do, and (b) it's often misconstrued as yet another test. But as James Popham notes, formative assessment is a test-supported *process* rather than a test (Popham, 2009).

The goal of this chapter is to describe formative assessment fully and also present a special approach to formative assessment called *stealth assessment*. Therefore, we discuss (a) measurement and assessment, (b) summative and formative assessment, and (c) formative and stealth assessment. Within each of these sections, we provide definitions, examples, and relevant research. We conclude this chapter with recommendations to help bring formative assessment into the classroom and a discussion about how stealth assessments fit well with a systematic approach to instructional design.

Measurement and Assessment

Different models of educational measurement are associated with different instructional practices in the classroom and thus have different effects on student learning. Historically, the main aim of measuring students' educational progress was to identify differences among students in order to rank order them by achievement. This type of measurement model makes heavy use of summative assessment, which is useful for accountability purposes but only marginally useful for guiding day-to-day instruction and supporting student learning. In contrast, student-centered measurement models rely mostly on formative assessment, which is associated with meaningful feedback that can be very useful in guiding instruction and supporting student learning.

Assessment is a general term that typically applies to individuals and may include testing, observation, and so forth. Progress toward educational goals is usually assessed through testing of some type. Assessment can refer to both an instrument and a process by which information is obtained relative to a known objective or goal (Shute, 2009). Since inferences are made about what a person knows on the basis of responses to a limited number of assessment tasks or items, there is always some uncertainty in inferences made on the basis of assessments. The goal in educational measurement is to minimize uncertainty or error; thus key aspects of assessment quality are validity and reliability. Reliability refers to the consistency of assessment results—the degree to which they rank order students in the same way. 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 task or test results to underlying competencies.

The focus of this chapter concerns not only measuring students' existing and emergent competencies accurately and reliably but also using that information to render diagnoses and instructional support. Consequently, the focus is on formative assessment (FA) rather than summative assessment. Later, we describe stealth assessment which involves embedding formative assessment into the learning environment such that it is invisible and hence does not disrupt learning and engagement.

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 choice and use of a particular type of assessment depend on the educational purpose. As mentioned earlier, schools in the United States today 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. That is, one current downside of the assessment-for-learning model is that it is often implemented in a non-standardized and hence less rigorous manner than summative assessment, and thus can hamper the validity and consistency of the assessment tools and data (Shute & Zapata-Rivera, 2010).

Summative Assessment

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

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 students' 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 or her students on some current event, immediately analyzing their scores, and then refocusing his or her lesson to straighten out a prevalent misconception shared by the majority of students in the class. This type of assessment is intended to be 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 (e.g., Black & Wiliam, 1998; Hindo, Rose, & Gomez, 2004; Schwartz, Bransford, & Sears, 2005). In addition to providing teachers with evidence about how their class is learning so that they can revise instruction appropriately, formative assessment directly involves students in the process, such as by providing feedback that will help them gain insight about how to improve, and by suggesting (or implementing) instructional adjustments based on assessment results.

While the scope of what comprises an assessment for formative purposes is quite broad (e.g., informal data, test responses, homework, observations), what is key in the definition is that the information or the evidence is used as feedback—by teachers (or systems) and students to improve teaching and learning, respectively. It is essential that an FA system includes support tools to help teachers learn to implement the full range of assessment types, gather evidence, make sense of the data, and adjust instruction accordingly. Such support tools may reside within a professional development strand related to the FA system. An FA system should also provide support for learners to help them improve motivation, volition, self-efficacy, problem-solving skills, and so on.

Finally, notice that we use the term "formative assessment" throughout the chapter as if it were a singular entity, but there are actually two different faces of FA which may be construed as residing at opposite ends of a continuum. That is, at one end of the continuum lives formal FA, which relates to the more traditional, teacher-centric view of formative assessment; this involves administering tasks and quizzes to students, gathering students' results, and then either providing feedback or altering instructional activities on the basis of the data. The other end of the continuum—informal FA—involves the student-centric, interactive classroom activities and discussions that occur, often spontaneously, in various learning environments. Both formal and informal FA provide evidence to teachers and students about learning progress.

Table 25.1 characterizes four assessment variables (main role in the classroom, frequency of administration, typical format, and feedback) that are characteristic of summative and formative assessment. The examples, per variable, for summative and formative assessment are illustrative and not exhaustive (e.g., formative assessment formats may include other types besides constructed response, such as oral response and even multiple-choice questions). Also note that neither type of assessment is an educational panacea—both have strengths and limitations. Table 25.1 is intended to convey general aspects of each approach in terms of the variables and should not be viewed as definitive categorizations.

Research on Formative Assessment in the Classroom

Research suggests that well-designed and implemented formative assessment is an effective strategy for enhancing student learning. Evidence to date suggests that students in classes where formative assessment was implemented learned in 6 months what would have taken a year in other

Table 25.1 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. These are typically conducted at the end of a major event (e.g., unit, marking period, school year)	Intermittent, formative assessment. The focus is 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 normally informal
Format of assessment	Objective assessments, often using selected responses. The focus is on whether the test is valid and reliable 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)

classes (Wiliam, 2006). Studies indicate that the regular use of classroom formative assessment could raise student achievement by 0.4–0.7 of a standard deviation (Black & Wiliam, 1998)—enough to catapult the United States into the top five countries in the international rankings for math achievement (Wiliam & Thompson, 2007). Finally, there is evidence that formative assessment can promote significant gains in student self-efficacy and motivation (Kim, 2007), which are predictors of high school graduation (Black, Harrison, Lee, Marshall, & Wiliam, 2003). Another important finding from studies on formative assessment relates to the benefits for disadvantaged and low-achieving students (e.g., Fuchs et al., 1997).

When teachers know how students are progressing and where they are having problems, they can use that information to make real-time instructional adjustments such as reteaching, trying alternative instructional approaches, altering the difficulty level of tasks or assignments, or offering more opportunities for practice. Again, FA in this sense has been shown to improve student outcomes and achievement (Black & Wiliam, 1998; Shute, Hansen, & Almond, 2008).

Feedback is an important and direct component of good FA, and should generally guide students toward obtaining their goals. Helpful feedback provides specific comments to students about errors and suggestions for improvement. It also encourages students to focus their attention thoughtfully on a specific task rather than on getting the right answer or a passing grade (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Shute, 2008). This may be considered a direct application of FA.

A more indirect way (compared to feedback) of helping students learn via formative assessment includes instructional adjustments that are based on assessment results (Stiggins, 2002). Different types of assessment data can be used by the teacher to support learning, such as diagnostic information relating to levels of student understanding, and readiness information indicating whether or not a student is ready to begin a new lesson or unit. Examples of instructional support include (a) recommendations about how to use assessment information to alter instruction (e.g., speed up, slow down, give concrete examples), and (b) suggestions for what to do next, links to Web-based resources, and so forth. However, there is much room for improvement in teachers' formative use of assessment results, as one of the most important aspects of formative use (responding to results by modifying instruction and identifying alternative pedagogies) is the least used by classroom teachers and the most neglected with respect to professional development (see Lai, 2009).

Research on Formative Assessment in Computer-Based Learning Environments

A growing number of computer-based educational systems are employing formative assessment as well. A good example of such systems is a Web-based formative assessment platform called ASSISTment (Feng, Heffernan, & Koedinger, 2006; Koedinger, McLaughlin, & Heffernan, 2010). ASSISTment is a Web-based platform that allows teachers to develop formative assessments for fourth- to tenth-grade mathematics classes. In a recent study, Koedinger and his colleagues (2010) reported that the schools using ASSISTment significantly outperformed matched schools on the state mathematics test.

Another example of a computer-based formative assessment system is ACED (Adaptive Content with Evidence-based Diagnosis) (Shute, Graf, & Hansen, 2005). This system uses an evidence-centered design approach (Mislevy, Steinberg, & Almond, 2003) to create an adaptive, diagnostic assessment system to assess and support pre-algebra knowledge and skills. Instructional support is in the form of elaborated feedback. A study was conducted examining its efficacy (Shute et al., 2008). The key issue was whether the inclusion of the feedback into the system (a) impairs the quality of the assessment (relative to validity, reliability, and efficiency), and (b) does, in fact, enhance student learning. Results from a controlled evaluation testing 268 ninth-grade

Table 25.2 Summary of key formative assessment features

Feature	Rationale
Improves student learning	A primary purpose of an FA system is to enhance (or support) student conceptual development as well as skill acquisition. Two kinds of data to support learning include (a) *diagnostic information* relating to levels of understanding and particular misconceptions where the information from diagnostic tasks should be instructionally tractable (i.e., neither too general nor too specific) and (b) *readiness information*, where a general FA task is administered at the outset of a class or a unit and results can show who, in the class, is ready (or not) to begin a new lesson or unit
Promotes student self-efficacy	Feedback in FA should generally guide students through toward obtaining their goal(s) (Ramaprasad, 1983; Sadler, 1989). The most helpful type of formative feedback (on tests, homework, and classroom activities) provides specific comments to students about errors, and specific suggestions for improvement, and encourages students to focus their attention thoughtfully on the task rather than on simply getting the right answer (Bangert-Drowns et al., 1991; Elawar & Corno, 1985; Shute, 2008). This type of feedback may be particularly helpful to lower achieving students because it emphasizes that students can improve as a result of effort rather than be doomed to low achievement due to some presumed lack of innate ability (e.g., Hoska, 1993)
Provides timely feedback	Feedback must be timely to be useful (e.g., Corbett & Anderson, 1989). Whenever possible, the FA system should provide immediate feedback (ideally immediately, but within "same day" time frame). Feedback can be directed to students (e.g., regarding performance on computer-based tasks) or teachers (e.g., summary reports on classroom performance)
Provides information at multiple levels of aggregation	FAs should report out *individual* data and may be *aggregated* to subgroup and full-group levels. Teachers and administrators may be able to specify subgroups based on student demographic variables (e.g., gender, race, attendance, mobility, socioeconomic status, etc.) and also use FA results to create groups with similar performance on specified tasks or sets of tasks
Provides low-to-mid stakes assessment	Given the relatively low-stakes and informal nature of FAs, they should mostly be of two levels: low and intermediate (not high-stakes). Higher degrees of standardization in FAs may occur in certain computer applications. Also note that "low-stakes" does not mean they will be low in reliability or validity (see Shute et al. (2008) for an example of a reliable and valid FA system)
Uses developmental models	Competency models should include developmental aspects that provide pre- and post-requisite relationship information. The function of the developmental part of the models relates to (1) *actual* learning (self- or criterion referenced), and (2) *potential* learning (forecasting near and far term potential—via Zone of Proximal Development and "end of school year" growth modeling research ideas)

students showed that the quality of the assessment was unimpaired by the provision of feedback. Moreover, students using the ACED system showed significantly greater learning of the content compared with a control group. These findings suggest that assessments in other settings (e.g., standardized, state-mandated tests) might be augmented to support student learning with instructional feedback without jeopardizing the primary purpose of the assessment.

Table 25.2 summarizes the key features of formative assessment, along with a brief discussion of each feature.

So far, we have focused on FA. But now consider the following. Rather than stopping an instructional episode at various times to collect information from students and provide support as warranted, what if there was a way to embed FA so deeply in the fabric of the learning environment that the distinction between learning and assessing became completely blurred? This idea, called stealth assessment, is presented next.

Formative and Stealth Assessment

New directions in educational and psychological measurement allow more accurate estimations of students' competencies, and new technologies permit us to administer formative assessments during the learning process; extract ongoing, multifaceted information from a learner; and react in immediate and helpful ways. When formative assessments are seamlessly woven into the learning environment and are thus invisible to learners, we call this *stealth assessment* (Shute, 2011; Shute, Ventura, Bauer, & Zapata-Rivera, 2009).

Stealth assessment can be accomplished via automated scoring and machine-based reasoning techniques to infer things that would be too hard or time consuming for humans (e.g., estimating values of evidence-based competencies across a network of skills). One big question is how to make sense of rich data collected in order to provide meaningful feedback and other support for learning. Another major question concerns the best way to communicate a variety of student-performance information in a way that can be used to inform instruction and enhance learning.

Definition of Stealth Assessment

Stealth assessment is an evidence-based approach to assessment where the tasks that students are engaged with are highly interactive and immersive, such as within video games or other computer-based instructional systems. Like FA, stealth assessment is intended to support learning of important content and key competencies. This represents a quiet-yet-powerful process by which learner performance data is

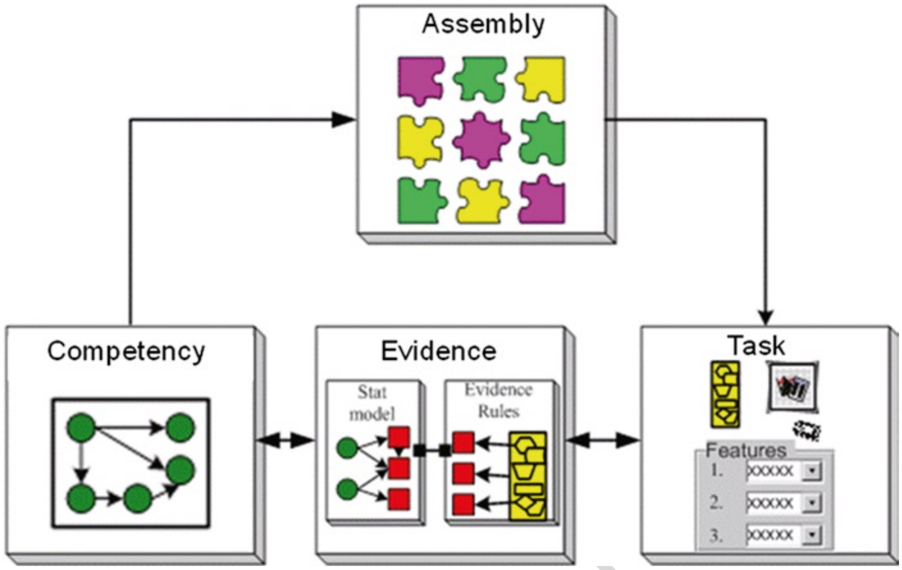

Fig. 25.1 Conceptual assessment framework of ECD (adapted from Mislevy et al., 2003)

continuously gathered during the course of playing/learning and inferences are made about the level of relevant competencies (see Shute et al., 2009). Inferences on competency states are stored in a dynamic model of the learner. Stealth assessment is intended to support learning and maintain flow, defined as a state of optimal experience where a person is so engaged in the activity at hand that self-consciousness disappears, sense of time is lost, and the person engages in complex, goal-directed activity not for external rewards, but simply for the exhilaration of doing (Csikszentmihalyi, 1990). Stealth assessment is also intended to remove (or seriously reduce) test anxiety while not sacrificing validity and reliability (Shute et al., 2008). Again, the goal is to blur the distinction between assessment and learning.

Key elements of the approach include (a) evidence-centered assessment design, which systematically analyzes the assessment argument concerning claims about the learner and the evidence that supports those claims (Mislevy et al., 2003), and (b) formative assessment and feedback to support learning (Black & Wiliam, 1998; Shute, 2008). Additionally, stealth assessment provides the basis for instructional decisions, such as the delivery of tailored content to learners (e.g., Shute & Towle, 2003; Shute & Zapata-Rivera, 2008). Information is maintained within a learner model and may include cognitive as well as noncognitive information comprising an accurate and up-to-date profile of the learner.

Evidence-centered assessment design (ECD), the key element for stealth assessment, is a conceptual design framework to help in the creation of coherent assessments. It supports a broad range of assessment types, from classroom quizzes to simulation-based assessments (Mislevy et al., 2003). The conceptual framework includes several models that work together to answer specific questions, such as "what attributes are to be measured?" and "how do we score them?" (see Fig. 25.1).

The competency model defines variables related to students' knowledge, skills, abilities, and other attributes that we wish to measure. This model accumulates and represents current beliefs about targeted aspects of skill, expressed as probability distributions per variable (Almond & Mislevy, 1999). The evidence model provides detailed instructions about (a) what the student says or does that can count as evidence for those skills (Steinberg & Gitomer, 1996), and (b) how the evidence statistically links to variables in the competency model (Mislevy, 1994). Task/action models express situations that can evoke required evidence. And the assembly model specifies how the competency, evidence, and task/action models work together to form a valid assessment.

Example of a Stealth Assessment

To illustrate the stealth assessment approach, here is an example relating to creative problem solving in a commercial game called Oblivion (*The Elder Scrolls IV: Oblivion,* 2006, by Bethesda Softworks). *Oblivion* is a first-person, 3D role-playing game that is set in an imaginary medieval world. Players enter the game by selecting a character to play (e.g., Argonian, Orc, or Dark Elf). Each character has a particular specialization (e.g., combat, stealth, and magic) and special abilities. The primary goal of the game is to develop the character's skills by completing a series of quests. These quests represent the character's journey to save the empire

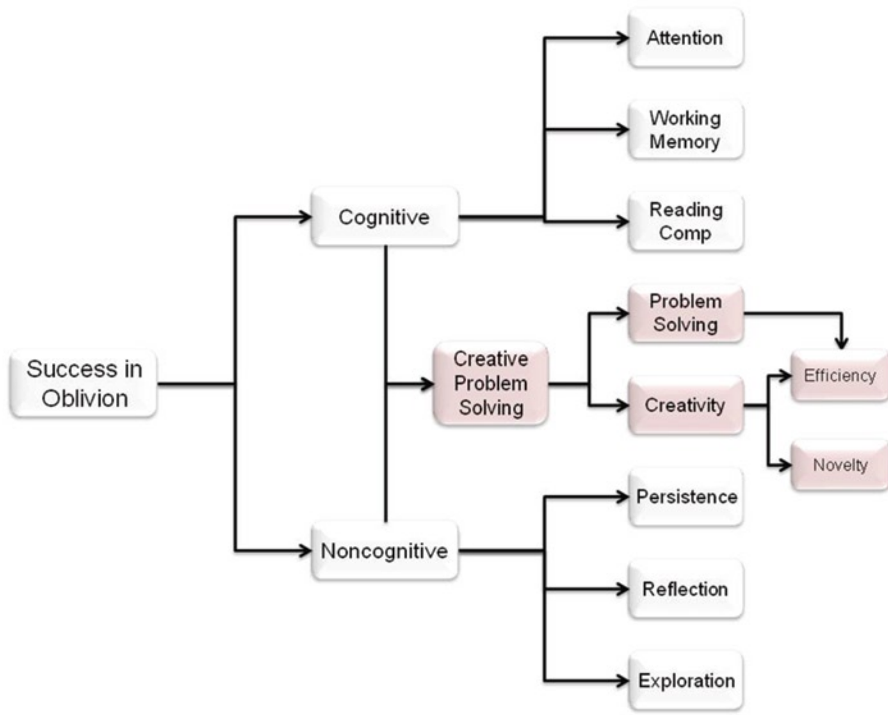

Fig. 25.2 Illustrative competency model for Oblivion (from Shute et al., 2009)

Table 25.3 Example of action model with indicators for novelty and efficiency

Action	Novelty	Efficiency
Swim across the river	$n=0.12$	$e=0.22$
Levitate over the river	$n=0.33$	$e=0.70$
Freeze the river with a spell and slide across	$n=0.76$	$e=0.80$
Find a bridge over the river	$n=0.66$	$e=0.24$
Dig a tunnel under the river	$n=0.78$	$e=0.20$

from dark magic, and are typically quite complex problems that players need to solve. During the course of the game, there are about 20 skills that a character needs to develop (e.g., alchemy, illusion, and heavy armor) to level up or to avoid being killed by dark monsters.

Creative problem solving is the main competency in the example, defined as the process of coming up with novel but efficient solutions to a given problem. The shaded competency model variables in Fig. 25.2 represent the nodes of interest in this example.

The evidence model links the specific actions that a player takes in the game with relevant competency variables. This requires the specification of particular observations, and how they differentially inform the level of mastery for different competency variables. The statistical machinery (such as IRT or Bayesian networks) serves to "glue" this information together (i.e., the observable performance data with the unobservable competency variables).

The action model (i.e., task model) in the example relates to the various quests and possible actions that players take in relation to quests. For example, consider a player faced with the problem of having to cross a river full of dangerous fish. Table 25.3 contains a list of actions to solve this problem, as well as the indicators that may be learned from real data, or elicited from experts. For the system to learn indicator values from real data, estimates of *novelty*, for example, may be defined in terms of the frequency of use across all players. For instance, swimming across the river is a high-frequency, common solution, thus associated with a low "novelty weight." An estimate of *efficiency* may be defined in terms of the probability of successfully solving a problem given a set of actions—based on time and resources expended.

Swimming across the river would thus have a low efficiency value because of the extra time needed to evade the dangerous fish. On the other hand, digging a tunnel under the river to get to the other side is judged as highly novel, but less efficient than, say, freezing the water and simply sliding across, the latter being highly novel and highly efficient. The indicator values shown in Table 25.3 were elicited from two *Oblivion* experts, and they range from 0 to 1. Higher numbers relate to greater levels of both novelty and efficiency.

Actions can be captured in real time as the player interacts with the game, and associated indicators can be used to provide evidence for the appropriate competencies. This is accomplished via the evidence model using Bayesian network software. Figure 25.3 shows a Bayes net after a player elected to cross the river by levitating over it.

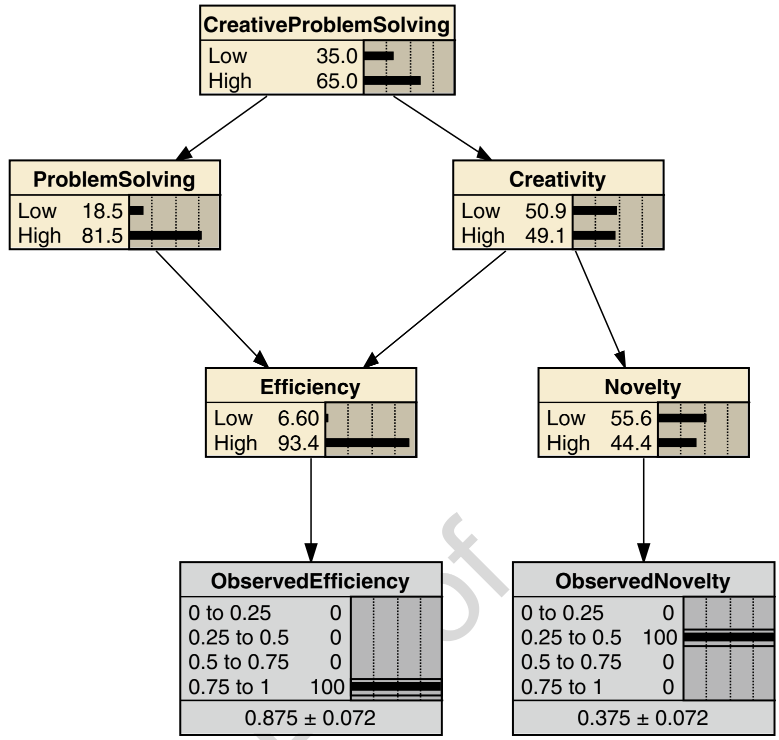

Fig. 25.3 Bayes net estimates from levitating over the river (from Shute et al., 2009)

Even though the player evidenced just average creativity in that solution, the parent node of creative problem solving infers that she is somewhat "high" on this attribute—illustrating that problem solving (based on efficiency) is a more valued competency than creativity, based on the way that the conditional probability distributions were set up in the competency model. Further, the player has more chances to improve this skill during game play. This information can be used in two different ways: (a) as formative feedback, which can be directly communicated to the learner, and (b) adjusting the sequence of quests to focus more emphasis on improving creativity.

Conclusion

In this chapter, we discussed formative assessment in relation to measurement and summative assessment. We also described stealth assessment as a particular instantiation of formative assessment, as employed within a video game or other immersive environment. Despite their intuitive appeal, both formative and stealth assessment have some challenges that need to be addressed for them to be widely adopted in classrooms today.

First, for formative assessment to be embraced more widely there should be more support—such as through professional development—for teachers. This would enable them to be more comfortable and skilled using formative assessment in their classrooms. In particular, teachers should learn to (a) diagnose students' competencies (at various grain sizes) based on different sources of information, (b) figure out what to do next given the obtained data, and (c) build up and employ a pool of rich tasks, probing questions, and other instructionally fruitful activities that can serve to elicit more evidence to inform student models and concurrently support students' learning. In short, teachers should acknowledge that formative assessment is intended to support their decision making for instructional adjustment to help all students grow and learn.

Following are ten recommendations for teachers about how to effectively use formative assessment in the classroom:

1. *Cognitive research*. Employ assessments that have been designed on a cognitive-developmental research foundation.
2. *Complex tasks*. Engage students in cognitively demanding tasks, i.e., ones that actually engage students in thinking about an issue or a problem.
3. *Learning goals*. Inform students clearly of the specific (and more general) learning goals being sought in the lesson or across longer units.
4. *Administration*. Administer assessments (of all types) frequently and usually informally, and require full-class participation in the ongoing, interactive dialog.

5. *Feedback*. Give feedback to students in the form of constructive comments, not grades.
6. *Personal accountability*. Provide students with opportunities to assess themselves and/or their peers to support personal accountability and autonomy.
7. *Evidence-based diagnosis*. Use evidence from formal and informal FAs as the basis for diagnosing students' progress (or lack thereof).
8. *Preplan questions and paths*. Plan questions in advance that probe students' understanding and craft alternative instructional paths based on response patterns.
9. *Leverage prior knowledge*. Build on students' preexisting knowledge and understanding—even if it requires going back through previously instructed material.
10. *Collaboration*. Meet regularly with other teachers to select and share good tasks, discuss student work, plan effective questions, discuss "lessons learned," and so on.

Implementing stealth assessment also poses its own set of challenges. The competency model, for example, must be developed at an appropriate level of granularity to be implemented in the assessment. Too large a grain size means less specific evidence is available to determine student competency, while too fine a grain size means a high level of complexity and increased resources to be devoted to the assessment. In addition, developing the evidence model can be rather difficult in a gaming environment when students collaborate on completing quests. For example, how would you trace the actions of each student and what he/she is thinking when the outcome is a combined effort? Another challenge comes from scoring qualitative products such as essays, student reflections, and online discussions where there remains a high level of subjectivity even when teachers are provided with comprehensive rubrics.

How do teachers fit into this effort? In games designed for educational purposes, the system can allow teachers to view their students' progress relative to the students' competency models. Teachers would then use that information as the basis for altering instruction or providing formative feedback. For example, if the competency models during a quest showed evidence of a widespread misconception, the teacher could turn that into a teachable moment, or may choose to assign struggling students to team up with more advanced students in their quests.

Information about students' competencies may also be used by the game system to select new gaming experiences (e.g., more challenging, ill-structured problems could be presented to students exhibiting high creative problem-solving skills). In addition, up-to-date estimates of students' competencies, based on assessment information handled by the statistical machinery (e.g., Bayes nets), can be integrated into the game and explicitly displayed as progress indicators. Players could then see how their competencies are changing based on their performance in the game. Most games already include status bars, representing the player's current levels of game-related variables. Imagine adding high-level competency bars that represent attributes like creative problem solving, persistence, and leadership skill. More detailed information could be accessed by clicking the bar to see current states of lower level variables. And like health status, if any competency bar gets too low, the student needs to act to somehow increase the value. Once students begin interacting with the bars, metacognitive processes may be enhanced by allowing the player to see game- or learning-related aspects of their state. Viewing their current competency levels and the underlying evidence gives students greater awareness of personal attributes. In the literature, these are called "open student models" and they have been shown to support knowledge awareness, reflection, and learning (Bull & Pain, 1995; Hartley & Mitrovic, 2002; Kay, 1998; Zapata-Rivera & Greer, 2004; Zapata-Rivera, Vanwinkle, Shute, Underwood, & Bauer, 2007).

How is stealth assessment related to the design of instructional systems? Gustafson and Branch (2002) describe five core elements of instructional design: analysis, design, development, implementation, and evaluation. These factors ensure coherence among instructional goals and strategies, as well as the effectiveness of the instruction. Moreover, these five elements should be used iteratively, and evaluation should reside at the center of the iterative revision process. Information obtained from any stealth assessment can also be used by instructional designers to improve learning/instructional systems. For example, information from a stealth assessment may show that many students had difficulty with a particular task. The instructional designer could then examine the task to see if revisions are warranted.

In addition, components of a stealth assessment (e.g., competency, evidence, and task models) are compatible with steps in the instructional design process such as task and content analysis and the development of performance measures. A common goal of both stealth assessment and instructional design is to coherently align learning objectives with how they are measured. Therefore, if instructional designers work closely with assessment developers to design and develop a learning system that has built-in stealth assessment, it can optimize the effectiveness of the instruction.

In conclusion, the ideas in this chapter relate to using formative assessment (in the classroom) and stealth assessment (in immersive learning environments). In both cases, this would help not only to collect valid evidence of students' competency states and support student learning but also to reduce teachers' workload in relation to managing the students' work products. This would allow teachers to focus their energies on the business of fostering student learning. The ideas in this chapter are intended to help teachers facilitate student learning, in a fun and engaging manner, of educationally valuable skills.

References

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 2*(3), 223–237.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. New York, NY: Open University Press.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Bull, S., & Pain, H. (1995). "Did I say what I think I said, and do you agree with me?": Inspecting and questioning the student model. In J. Greer (Ed.), *Proceedings of AI-ED'95—7th World Conference on Artificial Intelligence in Education* (pp. 501–508). Virginia: AACE.

Chappius, S., & Chappius, J. (2008). The best value in formative assessment. *Educational Leadership, 65*(5), 14–19.

Corbett, A. T., & Anderson, J. R. (1989). Feedback timing and student control in the LISP intelligent tutoring system. In D. Bierman, J. Brueker, & J. Sandberg (Eds.), *Proceedings of the Fourth International Conference on Artificial Intelligence and Education* (pp. 64–72). Amsterdam, The Netherlands: IOS Press.

Council of Chief State School Officers [CCSSO]. (2004). *Indicators of quality of teacher professional development and instructional change using data from surveys of enacted curriculum: Findings from NSF MSP-RETA project*. Washington, DC.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optical experience*. New York, NY: Harper Perennial.

Elawar, M., & Corno, L. (1985). A factorial experiment in teachers' written feedback on student homework: Changing teacher behavior a little rather than a lot. *Journal of Educational Psychology, 77*(2), 162–173.

Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006). *Addressing the testing challenge with a web-based e-assessment system that tutors as it assesses*. Paper presented at the 15th International Conference on World Wide Web, Edinburgh, Scotland.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Katzaroff, M., & Dutka, S. (1997). Effects of task-focused goals on low-achieving students with and without learning disabilities. *American Educational Research Journal, 34*(3), 513–543.

Gustafson, K. L., & Branch, R. M. (2002). What is instructional design? In R. A. Reiser & J. V. Dempsey (Eds.), *Trends and issues in instructional design and technology* (pp. 16–25). Columbus, OH: Merrill Prentice Hall.

[AU2] Hartley, D., & Mitrovic, A. (2002). Supporting learning by opening the student model. In S. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Proceedings 6th International Conference on Intelligent Tutoring Systems* (pp. 453–462). Biarritz, France.

[AU3] Hindo, C., Rose, K., & Gomez, L. M. (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* (pp. 606–609). Mahwah, NJ: Erlbaum.

Hoska, D. M. (1993). Motivating learners through CBI feedback: Developing a positive learner perspective. In V. Dempsey & G. C. Sales (Eds.), *Interactive instruction and feedback* (pp. 105–132). Englewood Cliffs, NJ: Educational Technology Publications.

Jennings, J., & Rentner, D. S. (2006). Ten big effects of the no child left behind act on public schools. *Phi Delta Kappan, 88*(2), 110–113.

Kay, J. (1998). *A scrutable user modelling shell for user-adapted interaction* (Doctoral thesis, University of Sydney, Sydney, Australia). Retrieved from http://sydney.edu.au/engineering/it/~judy/Homec/Pubs/thesis.pdf.

Kim, C. (2007). *Effects of motivation, volition and belief change strategies on attitudes, study habits and achievement in mathematics education*. Doctoral dissertation, Florida State University, Tallahassee, FL.

Koedinger, K., McLaughlin, E., & Heffernan, N. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research, 4*, 489–510.

Lai, E. R. (2009). *Interim assessment use in Iowa elementary schools* (Doctoral thesis, University of Iowa, Iowa City, USA). Retrieved from http://ir.uiowa.edu/etd/393/.

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

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

Popham, W. J. (2009). A process—Not a test. *Educational Leadership, 66*(7), 85–86.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioural Science, 28*(1), 4–13.

*Sadler, D. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144.

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

Shute, V. J. (2007). Tensions, trends, tools, and technologies: Time for an educational sea change. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 139–187). New York, NY: Lawrence Erlbaum Associates.

*Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. doi: 10.3102/0034654307313795.

Shute, V. J. (2009). Simply assessment. *International Journal of Learning and Media, 1*(2), 1–11. doi:10.1162/ijlm.2009.0014.

*Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age Publishers.

Shute, V. J., Graf, E. A., & Hansen, E. (2005). Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities. In L. PytlikZillig, R. Bruning, & M. Bodvarsson (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 169–202). Greenwich, CT: Information Age Publishing.

Shute, V. J., Hansen, E. G., & Almond, R. G. (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 in Education, 18*(4), 289–316.

Shute, V. J., & Towle, B. (2003). Adaptive e-learning. *Educational Psychologist, 38*(2), 105–114.

*Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Mahwah, NJ: Routledge, Taylor and Francis.

Shute, V. J., & Zapata-Rivera, D. (2008). Adaptive technologies. In J. M. Spector, D. Merrill, J. van Merriënboer, & M. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 277–294). New York, NY: Lawrence Erlbaum Associates.

Shute, V. J., & Zapata-Rivera, D. (2010). Intelligent systems. In E. Baker, P. Peterson, & B. McGaw (Eds.), *Third edition of the* *international encyclopedia of education* (pp. 75–80). Oxford, UK: Elsevier.

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

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758–765.

Symonds, K. W. (2004). *After the test: Closing the achievement gaps with data*. Naperville, IL: Learning Point Associates.

Wiliam, D. (2006). Does assessment hinder learning? Speech delivered at the ETS Europe Breakfast Salon. Retrieved from http://www.decs.sa.gov.au/adelaidehills/files/links/williams_speech.pdf.

*Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Zapata-Rivera, D., & Greer, J. E. (2004). Interacting with inspectable Bayesian models. *International Journal of Artificial Intelligence in Education, 14*, 127–163.

Zapata-Rivera, D., Vanwinkle, W., Shute, V. J., Underwood, J. S., & Bauer, M. (2007). English ABLE. In R. Luckin, K. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education—Building technology rich learning contexts that work* (pp. 323–330). Amsterdam, The Netherlands: IOS Press.

Author Queries

Chapter No.: 25 0001957772

Queries	Details Required	Author's Response
AU1	Reference citation Shute (2011) in Abstract has been expanded as per Springer style. Please check if appropriate.	
AU2	Please provide publisher name for the reference Hartley & Mitrovic (2002) if possible.	
AU3	Please provide editors name for Hindo, Rose, & Gomez (2004).	