

# You Can't Fatten A Hog by Weighing It – Or Can You? Evaluating an Assessment for Learning System Called ACED

**Valerie J. Shute**, *Educational Psychology and Learning Systems Department, Florida State University, Tallahassee, FL 32306 USA*  
*vshute@fsu.edu*

**Eric G. Hansen, Russell G. Almond**, *Research and Development Center, Educational Testing Service, Princeton, NJ 08541 USA*  
*{ehansen, ralmond}@ets.org*

**Abstract.** The purpose of the study described in this paper was to evaluate the efficacy of an assessment for learning system named ACED (Adaptive Content with Evidence-based Diagnosis). We used an evidence-centered design approach to create an adaptive, diagnostic assessment system which includes five main models: competency, evidence, task, presentation, and assembly. We also included instructional support in the form of elaborated feedback. The key issue we examined 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 high-school 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. state-mandated tests) might be augmented to support student learning with instructional feedback without jeopardizing the primary purpose of the assessment.

**Keywords.** Adaptive sequencing, assessment for learning, Bayesian network, formative assessment, mathematics

## INTRODUCTION

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 student learning. In contrast, student-centered measurement models rely mostly on formative assessment, also known as “assessment *for* learning,” which can be very useful in guiding instruction and supporting student learning, but may not be particularly reliable or valid. That is, one 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 reliability of the assessment tools and data (Shute & Zapata-Rivera, in press). In addition to the wide variety and non-standardization of methods for implementing assessment for learning, there are also problems with accurately modeling the variety of types and levels of student knowledge within multifaceted environments. This poses a number of

psychometric challenges (e.g. modeling multiple abilities and other learner characteristics) regardless of the measurement model employed.

In conjunction with the measurement issues noted above, another problem we address in this research has a more human face. That is, teachers at all levels (primary, secondary, and higher education) have a very hard job. They need to not only instruct important concepts and skills but also assess students' competency levels in relation to important criteria (e.g. state or national standards). Furthermore, their time is usually quite limited, so the question is how to best juggle these activities of instruction and assessment. Too often, the result is "teaching to the test" where the test may be highly reliable, but limited in scope and depth which may seriously compromise the validity of score-based inferences (e.g. Popham, 2001; Shute, 2007). An alternative solution is to merge instruction and assessment within a single system. This "merging" idea is not new. For example, more than 15 years ago, Snow and Mandinach (1991) called for the development of principles for creating valid and useful instructional-assessment systems. Today, approaches that integrate assessment and instruction are emerging and go by a number of different names, such as: (a) "assessments for learning" (e.g. Black & Wiliam, 1998a, 1998b; Stiggins, 2002, 2006); (b) "informative assessments" (e.g. Bass & Glaser, 2004); (c) "educative assessments" (Wiggins, 1998); and (d) "assistsments" (e.g. Razzaq, Feng, Nuzzo-Jones, Heffernan, Koedinger, Junker et al., 2005). Each seeks to provide instruction with assessment that is sufficiently valid and reliable for its intended purpose. Controlled evaluations on particular components of such systems, however, are lacking.

The goal of the research reported in this paper was to design, develop, and evaluate an assessment tool that can also support student learning. The main issue we explore concerns whether combining assessment and instructional technologies into one system harms (or not) their original, respective purposes. For instance, does combining an adaptive test with feedback reduce its reliability, validity, or efficiency? Does choosing problems to rapidly converge on an accurate assessment impair students' learning? Towards that end, we describe our melded system and present the results from an evaluation of it in terms of how well it works: (a) as an assessment tool, and (b) to support learning. By analogy, can we actually "fatten a hog" (enhance student learning) with the same tool used to weigh (assess) it?

Our assessment for learning system is called ACED (Adaptive Content with Evidence-based Diagnosis). It assesses students' knowledge and skill levels of Algebra I content and combines adaptive task sequencing with instructional feedback to support student learning. Three main features of ACED were tested: feedback type, task sequencing, and competency estimation. The specific research questions we examine in this paper include the following (where the first two focus on learning and the third relates to assessment, but all three are intertwined):

- *Question 1:* Is elaborated feedback (i.e. task-level feedback that provides both verification and explanation for incorrect responses) more effective for student learning than simple feedback (verification only)?
- *Question 2:* Does adaptive sequencing of tasks for assessment have any negative impact on student learning?
- *Question 3:* Does the provision of task-level feedback render the assessment less valid, reliable, and/or efficient?

We begin with a brief review of the research relating to assessment for learning, task-level feedback, and adaptive sequencing of tasks, followed by an overview of the program (including its underlying models), and then the results from our evaluation of the system in relation to learning support and assessment quality.

## BACKGROUND AND RELATED PROJECTS

### Assessment for Learning

*If we think of our children as plants... summative assessment of the plants is the process of simply measuring them. The measurements might be interesting to compare and analyze, but, in themselves, they do not affect the growth of the plants. On the other hand, formative assessment is the garden equivalent of feeding and watering the plants - directly affecting their growth.”*  
Clarke (2001, p. 2).

An assessment *for* learning (AfL) approach to education involves weaving assessments directly into the fabric of the classroom or curriculum. Teachers or computer-based instructional systems<sup>1</sup> use the results from students' activities as the basis on which to adjust instruction to promote learning and understanding in a timely manner (Birenbaum et al., 2006; Leahy, Lyon, Thompson, & Wiliam, 2005; McTighe & O'Connor, 2005; Schwarz & Sykes, 2004; Shepard, 2003). This type of assessment is administered more frequently than summative assessment (Symonds, 2004), 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, 1998b). Unfortunately, while assessment *of* learning is currently well entrenched in most educational systems, assessment *for* learning is not.

In addition to providing teachers or automated instructional systems with evidence about how their students are learning so that they can revise instruction appropriately, AfL systems are intended to support student learning, such as by providing feedback that will help students gain insight about how to improve (Shute, 2008), and by suggesting or implementing instructional adjustments based on assessment results (Brookhart, 2001; Harlen, 2005; Peat & Franklin, 2002; Stiggins, 2002; Taras, 2002; Tiknaz & Sutton, 2006; Wiliam et al., 2004). These promises, however, need controlled evaluations to determine which features are most effective in supporting learning in a range of settings (see Jameson, 2007). Among the AfL features that have the greatest potential for supporting quality assessment and student learning, and which would be suitable for investigation are task-level feedback, and adaptive sequencing of tasks, which are described next.

### Task-level Feedback

Task-level feedback appears right after a student has finished solving a problem or task, and may be contrasted with (a) more general summary feedback, which follows the completion of the entire assessment, and (b) more specific step-level feedback which may occur within a task, as with intelligent tutoring systems (VanLehn, 2006). Task-level feedback typically provides specific and timely (often real-time) information to the student about a particular response to a problem or task, and may additionally take into account the student's current understanding and ability level. In general, feedback used in educational contexts is regarded as crucial to knowledge and skill acquisition (e.g. Azevedo & Bernard, 1995; Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Moreno, 2004), and may also influence motivation (e.g. Lepper & Chabay, 1985; Narciss & Huth, 2004). However, there are many different types of feedback, and – as described in Kluger and DeNisi (1996) and more recently

---

<sup>1</sup> Note that student models, as used within intelligent tutoring systems, represent a potential component of AfL. This will be illustrated in the Project Design and Development section of this paper.

in Shute (2008) – the literature on the type of feedback and the effects on student learning is often conflicted.

Immediate feedback on students' solutions to individual tasks has generally been shown to support student learning (e.g. Corbett & Anderson, 2001), especially when a response or solution is wrong. That is, when a student solves a problem correctly, it usually suffices to simply provide verification of the accuracy of the response (e.g. "You are correct"). But in the case of incorrect answers, some research has shown that it is more beneficial to provide not only verification of the incorrectness but also to provide an explanation of how one would determine the correct answer (e.g. Kluger & DeNisi, 1996; Mason & Bruning, 2001; Narciss & Huth, 2004). In this research, we focused on task-level feedback for incorrect answers and evaluated the contribution to learning that such elaborated (or conceptual) feedback provides relative to simple verification ("correct" or "incorrect") feedback.

### **Adaptive Sequencing of Tasks**

Adaptive sequencing of tasks within an assessment contrasts with the more typical fixed, linear sequencing of tasks or items. Adaptive sequencing usually entails making adjustments to the sequence of tasks based on determinations such as: (a) which task would be most informative for refining an estimate of the student's competency level, and (b) which task would be most helpful in supporting the student's progress to a higher competency level (see Shute & Zapata-Rivera, 2008). The rationale for employing an adaptive assessment is that students come to any new learning task with differing profiles. As educators, we want to take what we already know about students and add to that an understanding of what they are doing in real time in the AfL environment. We can then combine that information with knowledge about strategies for bringing individuals to a higher level of knowledge, and adapt content to carry out those strategies. According to Bass and Glaser (2004), taking full advantage of such assessments requires the use of adaptive techniques that yield information about the student's learning process and outcomes. This allows teachers or computer programs to take appropriate instructional actions and make meaningful modifications to instruction.

The ACED program uses only the first type of adaptivity (i.e. adaptivity for information gain within an assessment). Although this type of adaptivity does not necessarily have any optimality properties for promoting student learning, its adaptive algorithm tends to select tasks for which the student has an approximately 50-50 chance of solving correctly. These tasks are likely to reside within the student's zone of proximal development (Vygotsky, 1967) and hence may be good candidates for promoting learning, particularly if accompanied by feedback.

Most adaptive sequencing algorithms, like the one employed in ACED, work by maximizing a *quasi-utility*<sup>2</sup> to find the task that provides the most information about a student. Van der Linden (1998) catalogs many approaches; the most popular ones include: Fisher Information (used in IRT-CAT – item-response theory and computer-adaptive testing; Wainer, 1990; Wainer et al., 2000), minimum posterior variance (Owen, 1975; used in SIETTE, Conejo et al., 2004), and information gain (Millán & Pérez-de-la-Cruz, 2002). Most of these quasi-utilities assume that the competency domain

---

<sup>2</sup> A quasi-utility is used instead of a true utility to calculate the value of information because the process of eliciting a true utility is quite complex. The utility for the information from the item is related to the relative value of the various competency states (a complex elicitation problem). Almond (2007a) discusses the value of information for such decisions in more detail.

is unidimensional; however, Hensen and Douglas (2005) suggest using minimizing entropy for multivariate domains. Most choices of quasi-utility yield similar task sequences, and the common practice of matching the task difficulty to the current competency estimate is heuristic shorthand for choice of a task that is usually optimal with respect to the information gain criteria.

Madigan and Almond (1996) describe a property of applying these information optimization selection algorithms in multidimensional spaces that causes them to behave differently from the sequence an instructor would likely select. In particular, the algorithms tend to produce jumps, from topic to topic. For example, a full version of the competency model focused on in this paper (described in the next section) contained three branches: one each for arithmetic, geometric and other recursive sequences. The overall *Sequences* competency (i.e. the parent node) relies on information from all three branches. Consider what happens when the student solves (or fails to solve) an arithmetic sequence task. As the information about *Arithmetic Sequences* has just increased, the optimal information gain will come from one of the other two branches. Thus, the system would flit between the three kinds of sequences in a way that constantly forces the student to switch contexts (or cognitive gears). To get around this flighty behavior, Madigan and Almond proposed employing the critiquing strategy of Barr and Feigenbaum (1981). The idea is to suggest a *hypothesis* (e.g. the student is at least at the *medium* level of the *Arithmetic Sequences* competency) and test that hypothesis until sufficient information is obtained (i.e. the probability of it being true or false exceeds a threshold). The system would then move on to the next hypothesis (e.g. one related to *Geometric Sequences*). Madigan and Almond further suggest using the *expected weight of evidence* (EWOE; Good & Card, 1971) as the quasi-utility, in part because EWOE is always defined with respect to a hypothesis. Using this scheme, the instructor could have considerable freedom in choosing a series of hypotheses and desired confidence levels for changing to the next topic that can structure the content to meet a variety of classroom or student needs.

## PROJECT DESIGN AND DEVELOPMENT

Good assessments are key to obtaining relevant information on which to make valid inferences about students' knowledge and skill states. Moreover, accurate inferences of current knowledge and skill states support adaptive decisions that can promote learning. Assessment quality and learning effectiveness are thus the dual foci of this study. Consequently we conducted a controlled evaluation to examine the impact of task-level feedback and task sequencing on learning. We also determined the reliability and validity of the assessment information of our ACED system.

The topic of "sequences" (e.g. arithmetic, geometric, and other recursive sequences, such as the Fibonacci series) was selected for implementation based on interviews with school teachers in New Jersey, review of the National Council of Teachers of Mathematics (NCTM) standards, and state standards in mathematics. Details on the system's features and functionality may be found in Shute, Graf and Hansen (2005). This paper focuses on the results of our evaluation efforts, which involved testing the system in relation to its assessment quality and support of student learning. This evaluation of ACED utilized a subset of the overall competency model – *Geometric Sequences* (i.e. successive numbers linked by a common ratio). This set of competencies, about one third of the full set, is sufficiently rich yet not unwieldy to satisfy the constraints for a two-hour testing session. Figure 1 illustrates the main concepts (or "nodes") in the competency model, and the Appendix contains a detailed description of the competencies.

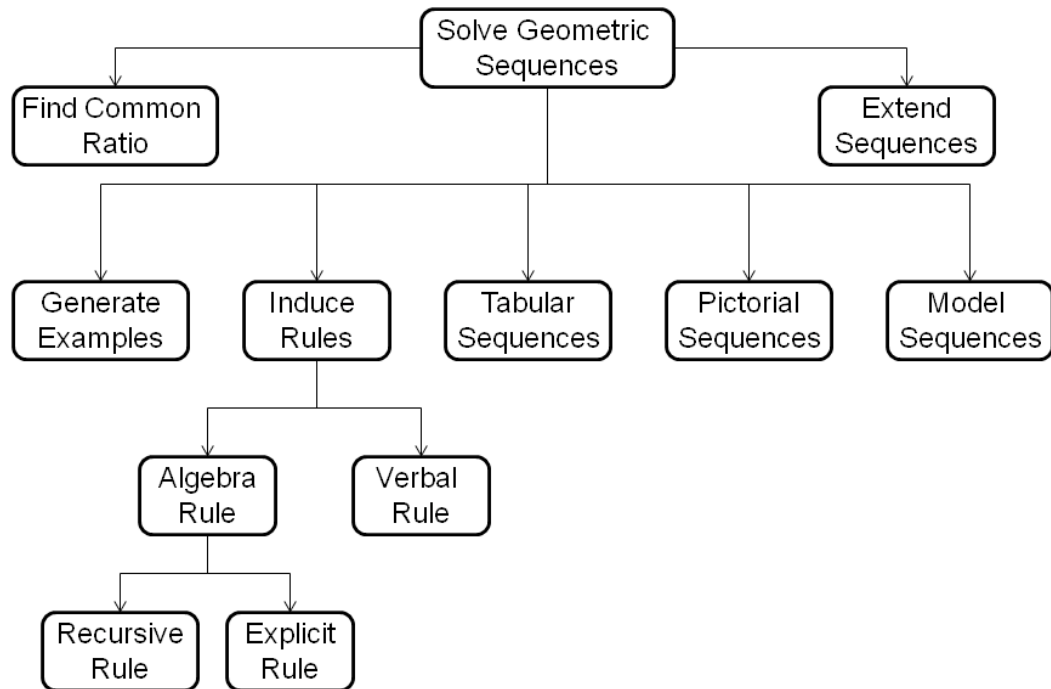


Fig. 1. ACED competencies in relation to Geometric Sequences.

## Assessment Design

We employed an evidence-centered design (ECD) approach (Mislevy, Steinberg, & Almond, 2003) to create the ACED system. ECD seeks to provide a coherent argument for the quality of an assessment. For example, it provides a framework for developing assessment tasks that elicit evidence (scores) that bears directly on the claims that one wants to make about what a student knows and can do. It is sometimes said that ECD simply formalizes the kind of thinking ordinarily done by expert assessment developers. However, by making this kind of thinking more widely available through reusable templates (task shells, design patterns, construct definition worksheets, etc.), the various participants in the design process – such as educators, assessment developers, measurement specialists, computer programmers – can communicate more effectively about key design issues, thereby more fully leveraging the diverse expertise of team members. This advantage may be especially important for assessment designs that are innovative in one way or another, such as the present project, which involves the integration of assessment and instruction. Thus, the basic idea of ECD is to specify the evidentiary argument of an assessment, including claims, evidence, and supporting rationales. By making the evidentiary argument explicit, the argument becomes easier to examine, share, and refine. Argument structures encompass, among other things, the claims one wants to make about what a student knows and can do and evidence to support those claims.

Figure 2 shows the basic structures of an evidence-centered approach to assessment design (Mislevy, Steinberg, & Almond, 2003).

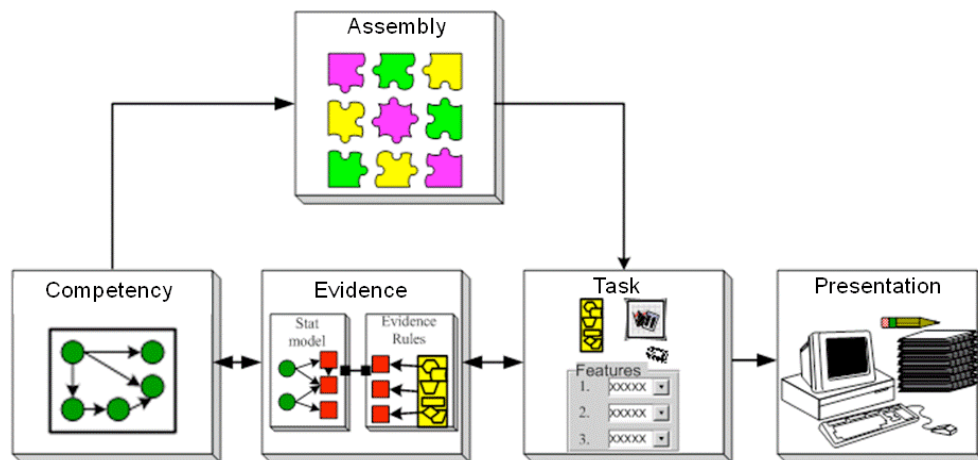


Fig.2. Central models of an evidence-centered assessment design.

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:

- What complex of knowledge, skills, or other attributes should be assessed?* A given assessment is meant to support inferences for some purpose, such as grading, promotion, providing diagnostic feedback, and so on. The competency model (CM; sometimes called the student model or proficiency model), describes knowledge, skills, and abilities about which inferences are intended (e.g. “This student has a high level of knowledge of solving geometric sequences”). In Figure 2, the various aspects of the targeted competency are represented by the circles. For the ACED system, the circle in the upper left of the CM box could be likened to the “Solve Geometric Sequences” node and the other circles to the other nodes shown in Figure 1. Values in the competency model express the assessor’s current belief about a student’s levels of knowledge, skill, and understanding of some content.
- What behaviors or performances should reveal those constructs?* The evidence model provides mechanisms for allowing student performances to serve as evidence for CM variables. The evidence model consists of two main parts: (a) the evidence rules (sometimes called the scoring model) and (b) the statistical model (sometimes called the measurement model). First, the evidence rules take as input the work product (represented by the rectangle with the jumble of shapes inside) resulting from the student’s interaction with a task. Depending on the type of task (item), the work product might be, for example, the selection of an answer option, a short answer, a piece of artwork for an art assessment, and so on. As output, the evidence rules produce observable variables (i.e. scores) that are evaluative summaries of the work products. For example, a very simple evidence rule might be, “If the student has selected option D, then the score takes the value of *correct*; otherwise the score takes the value of *incorrect*.” In this case, the evidence rule takes the student’s selection and the work product to produce a score of ‘correct’ or ‘incorrect.’ Second, the statistical model expresses the relationship, in probability or logic, between the competency model variables and the observable variables (scores). It enables updating the competency model variables in a way that aggregates scores across tasks or performances. The statistical model may be as simple as number-right scoring for a single competency variable. As will be explained later in

more detail, the ACED system employs Bayes net software to update an overall competency (Solve Geometric Sequences) and other competency variables. Bayes nets enable use of a statistical model that is far more sophisticated than simple number-right scoring. Note that the observable variables are the boxes at the center of the evidence model, and are used as the *output* from the evidence rules and the *input* to the statistical model.

- *What tasks should elicit those behaviors?* A task model (*TM*) provides a framework for characterizing and constructing situations with which a student will interact to provide evidence about targeted competencies. At a minimum it must specify (a) the kinds of material presented to the student as stimuli (e.g. text, tables, graphs, pictures) suggested by the picture and videotape icon in Figure 2; and (b) the kinds of work products produced by the student (e.g. selection, short answer, completed table, graph), represented in Figure 2 by the rectangle with the jumble of shapes inside of it. Task-model variables (or “features” in Figure 2) describe the characteristics of situations that will be controlled by the designer (Mislevy, Steinberg, & Almond, 2002). Task model variables play multiple roles, such as: (a) determining the evidentiary focus of a task (e.g. arithmetic, geometric or other recursive sequence; has table, graph, or picture; requires extending the series), (b) denoting the indicators of difficulty (e.g. series contains positive integers, positive and negative integers, real numbers), and (c) describing the incidental features that can be varied to make task variants (e.g. specific numeric values, names of objects in work problems). The value that a task model variable assumes in a particular task is determined by the characteristics of the presentation material used in that task.<sup>3</sup> 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 (e.g. scores, which are observable) about competencies (which are unobservable).
- *How are the tasks presented?* The goal of the presentation model is to de-couple the specification of the measurement properties of the task model from the way it is displayed in a particular environment. It is extremely useful for assessments that must be delivered in both paper and pencil and on-screen environments. In ACED it was used to design an alternative delivery mechanism using a “talking tactile tablet” for students with limited visual acuity<sup>4</sup> (for more, see Hansen & Shute, 2007; Shute, Hansen & Almond, 2007).
- *How much evidence must be collected?* The assembly model controls the sequence of tasks presented to the student and the rules for stopping the assessment. It controls both the

---

<sup>3</sup> This can be done either retrospectively (i.e. by analyzing the task and setting the variables appropriately), or prospectively (e.g. by creating materials to match a task with specific values). ACED used the task models in both of these ways, creating some tasks by reverse engineering existing tasks, and others via automatic item generation (Shute et al., 2005).

<sup>4</sup> A key to adapting assessments for individuals with special needs is to recognize that there are knowledge, skills and abilities (KSAs) that may be necessary to successfully complete the task that are not the focus of instruction. For example, solving an ACED problem with a picture requires the ability to visually distinguish objects in the picture. Most students in the target population have no difficulty with this skill, so this non-focal KSA is often overlooked. However, recognizing its influence is critical to adapting the assessment for individuals with low visual acuity. For ACED, in many cases the same task model could be used in both versions; in some cases, new task models needed to be developed to produce tasks that reduced demands on key non-focal KSAs. See Hansen and Mislevy (2005) and Hansen et al. (2005) for more details.



reliability of the assessment (largely related to test length) and the effective definition of the competencies. The definition of competencies is important because due to time considerations, the competencies in the competency model are usually defined at a fairly high grain size, and to properly cover the concept the student must address several different problem types. For example, if a competency called for *knowledge of graphs and tables*, then the assessment must cover both graph tasks and table tasks or else the effective meaning of that competency will change. The ACED assembly model called for the ability to deliver the assessment in two different ways: (a) a fixed (linear) sequence of tasks, and (b) using the weight of evidence algorithm (below) to adaptively select the task that would provide the most information about a student in terms of a target hypothesis.

It is useful to note that during assessment design we reason mostly in a deductive direction, from left to right in Figure 2. We attempt, for example, to deduce the kinds of scores (and performances) that would occur with a person who we know possesses a high (or low) level of the targeted competency. Expert judgments are critical in this deduction from (a) levels in the targeted competency to (b) expected scores as well as to (c) the task situations that would be needed to elicit those scores. The ACED project drew upon judgments of mathematics educators and researchers, as documented in Shute et al. (2005), to make this deductive argument.

On the other hand, during operational assessment, we reason primarily in an inductive direction, from right to left. For example, having obtained scores (based on performances prompted by the task situations), we attempt to infer the person's level (e.g. high or low) in terms of a targeted competency. In other words, during assessment use, the task features prompt the student to respond, thereby generating work products (e.g. selection of choices, short open-ended responses). The evidence rules are applied to the work products to produce scores and then the statistical model takes the scores (typically from multiple tasks) and updates the competency model.

During use of the ACED system, student responses are scored and competency model variables (estimates of learners' abilities) are updated. Different task models produce different tasks, which can vary along a number of dimensions (e.g. media type and difficulty level). Each node in the competency model is linked to at least six tasks, with three levels of difficulty (easy, medium, and hard) and two parallel tasks per level. Following are three levels of difficulty relating to the competency model variable: "*Find the common ratio in a geometric sequence.*"

EASY – *Find the common ratio for the following geometric sequence:*

**1, 2, 4, 8, 16, . . .** Enter your answer here \_\_\_\_\_

INTERMEDIATE – *Find the common ratio for the following geometric sequence:*

**4, -12, 36, -108, 324, . . .** Enter your answer here \_\_\_\_\_

DIFFICULT – *Find the common ratio for the following geometric sequence:*

**1.92, 0.96, 0.48, 0.24, 0.12, . . .** Enter your answer here \_\_\_\_\_

All of the ECD-based models in ACED were developed during the first year of the three-year project, while the second year focused on building the system's infrastructure, algorithms, tasks, and so on (see Shute et al., 2005 for details). The third year focused on evaluating the system, as described in this paper. While specifically evaluating the use of ECD for the ACED project is beyond the scope of this paper, it seems worthwhile to suggest that the use of ECD may have substantially contributed to the benefits that might accrue from ACED or derivative systems.

## Estimation of Knowledge and Skills

To estimate students' knowledge and skills from a set of observed outcomes from ACED tasks, we used a Bayesian network version of the Almond and Mislevy (1999) algorithm, which involved expressing the competency model and evidence models as Bayesian networks. Almond (2007b) describes the general approach to quantifying the networks. To build the competency model, one of our math content experts described the relationship between each connected pair of competency nodes in Figure 1 as a regression model, supplying the slope and correlation. To build the evidence models, we used a  $Q$ -matrix (Tatsuoka, 1983, 1985) describing the relationships between the tasks and targeted competencies, as well as the task developer's difficulty target to select parameters for a discrete IRT model (for more, see Almond et al., 2007).

Bayesian networks were chosen for several reasons, one of which was the ready availability of relevant software (i.e. StatShop software; Almond et al., 2006) to estimate students' competency levels incrementally for use with the adaptive assessment. Bayesian networks provide a flexible class of models that can readily incorporate prior information about the relationship of the competency variables of the kind described in Figure 1 (Almond et al., 2007). Multidimensional IRT (MIRT) would have provided an alternative framework to handle multiple competencies, except that implementing the Almond and Mislevy (1999) algorithm in that framework involves approximate solutions for high-dimensional integrals. Because ACED was only concerned with three levels of competency, Bayes nets are a decent approximation to the MIRT version of the algorithm.<sup>5</sup>

How does ACED compare with similar systems that employ a Bayesian model approach for estimating student competencies? Both ACED and a system called English ABLE (e.g. Zapata-Rivera et al., 2007) follow ECD principles to create the internal assessment models. They both also elicit initial prior and conditional probabilities from experts. They differ in that English ABLE makes use of IRT item parameters (e.g. item difficulty) from available TOEFL (Test of English as a Foreign Language) data to determine the conditional probability tables needed to drive the Bayesian network, while in ACED, item difficulty parameters were set to the fixed values of -1, 0, or 1 depending on whether our two math experts thought the task would be easy, medium difficulty, or hard. Also in English ABLE the current competency estimates were visible<sup>6</sup> to the student as they worked with the system, while ACED did not display scores until the session was complete. The structure of the Bayesian model described by Conati, Gertner and VanLehn (2002) represents both declarative and procedural knowledge, organized as rules and corresponding actions, while the nodes are divided into goals, facts, rule-application, and strategies. ACED competencies similarly represent declarative and procedural types of knowledge and skill, but the assessment model does not explicitly make use of those distinctions for instructional purposes (i.e. all knowledge types receive elaborated feedback to support learning).

---

<sup>5</sup> Although ECD was strongly influenced by the approach to model building described in Shafer (1976), we only considered models that worked within the Bayesian subset of the Dempster-Shafer (DS) theory; working with the full DS model would have added computational complexity, without providing much additional insight (Almond, 1995).

<sup>6</sup> Actually, the competency estimates were *indirectly* visible to the students, as they were ascribed to a character that the student was teaching. This allowed the score-based feedback to be decoupled to a certain extent from the student's self-concept (Zapata-Rivera et al., 2007).

## Adaptive Algorithm

We integrated an adaptive algorithm into ACED to determine which task to present next to maximize measurement accuracy. This involved calculating the expected weight of evidence (Good & Card, 1971; Madigan & Almond, 1996) per task as the basis for selecting the subsequent task. For example, suppose we hypothesize that a given student's ability to *Solve Geometric Sequences* (i.e. the parent node in the model) is at or above the *medium* level. The Bayes net can calculate the probability – expressed as the log odds – that this hypothesis holds. Now, suppose we observe a student's outcome from attempting to solve an ACED task. Entering this evidence into the Bayes net and updating the model will result in a change in the probabilities. The change in log odds is called the *weight of evidence* for a given hypothesis provided by evidence (Good, 1985). Typically, conclusions about a hypothesis will be based on the accumulation of many bits of evidence. Thus, weight of evidence can be used as the basis of graphical “explanations” of a hypothesis (Madigan, Mosurski, & Almond, 1997). For outcomes from tasks that have not been observed, one can calculate the *expected weight of evidence* (EWOE) under the hypothesis (Good & Card, 1971) defined as:

$$EWOE(H : T) = \sum_{j=1}^n \log \left[ \frac{P(t_j | h)}{P(t_j | \bar{h})} \right] P(t_j | h) \quad (1)$$

Here,  $T$  refers to task performance and  $H$  refers to the main hypothesis; either the main hypothesis is true ( $h$ ) or the alternative hypothesis is true ( $\bar{h}$ ). The variable  $n$  refers to the number of possible outcomes for each task. In ACED, there are two possible outcomes per task: correct or incorrect. The variable  $j$  represents the outcome index for a particular task, and the variable  $t_j$  is the value of the outcome (e.g. “correct”). The task that maximizes the expected weight of evidence will be chosen that is, in some sense, most informative about the hypothesis. This approach is related to the procedure commonly used in computer-adaptive testing of maximizing the Fisher information (Wainer et al., 2000). In ACED, computing the expected weight of evidence after each task response produces a new ordering of remaining tasks, unique for each student and based on the student's particular solution history.

Figure 3 illustrates the first 11 (of 63) adaptively-sequenced tasks for an actual student in relation to just one node (i.e. the main one – Solve Geometric Sequences) in the competency model. The prior probabilities assigned to this node at the outset are: High = 0.18, Medium = 0.26, and Low = 0.56. In this example, the first task (Identify Geometric Sequence) is selected and presented. Its difficulty level is “medium” (denoted in the task name, where 1=easy, 2=medium, and 3=hard). The student solves it incorrectly. At this point, the data in the student model are updated via the Bayes net in relation to all nodes. Notice the estimate for “High” competency for this specific node decreases slightly from the prior probability. ACED runs its adaptive algorithm on the remaining set of tasks and chooses the next task with the highest expected weight of evidence – in this case it is the same task type (i.e. Identify Geometric Sequence) but with a difficulty level of “hard” (denoted by the 3 in its name). Also note that this item selection is unlikely to be what a good teacher may choose for a student who failed to solve a medium-level item; a good teacher would tend to present an easier item. However, the student actually solves this hard task correctly, producing an increase in the estimates for “high” and “medium” competency, along with a decrease in the estimate that the student is “low” in relation to the particular node. Another hard item (of the same type) is presented, and the student solves that one correctly as well, producing another large increase in the probability that he is “high” in relation to the node. At the end of the first 11 items, the current (as of that point in time) estimate that the student is

“high” in relation to the Solve Geometric Sequences competency is 0.91 – suggesting that the student really understands this subject matter.

Node: *Solve Geometric Sequences*

Prior Probabilities:		0.18	0.26	0.56
Task	Score	High	Med	Low
IdentifyGeometricSequence2a	0	0.16	0.26	0.58
IdentifyGeometricSequence3a	1	0.35	0.35	0.30
IdentifyGeometricSequence3b	1	0.64	0.29	0.07
Identify GeometricSequence2b	1	0.83	0.16	0.01
VisualExtendTable2a	1	0.89	0.10	0.01
SolveGeometricProblem1a	0	0.78	0.21	0.01
SolveGeometricProblem1b	1	0.82	0.18	0.00
VisualExtendVerbalRule2a	1	0.85	0.15	0.00
ModelExtendTableGeometric3a	1	0.90	0.10	0.00
ExamplesGeometric2a	0	0.87	0.13	0.00
VisualExplicitVerbalRule3a	1	0.91	0.09	0.00
	⋮			

Fig.3. Example of a student’s adaptive sequence of tasks in the ACED system (first 11 of 63 tasks).

### Authoring Tasks

A total of 174 tasks (i.e. ACED items) were authored for the full topic of sequences, including arithmetic, geometric, and other recursive sequences, and a subset of 63 tasks comprised the geometric sequences branch of the competency model – the focus of this paper. Each task was linked, via the evidence model, to relevant competencies. A little over half of the items were multiple-choice format, with 4 to 5 options from which to choose. The remaining items required the student to enter a short constructed response (a number or series of numbers). Task development entailed not only writing the items, but also crafting feedback for answer choices related to multiple choice items and for likely common answers to constructed response items. Feedback for correct answers provided response verification (e.g. “You are correct!”). If the response was incorrect, the feedback was available in two forms: simple (e.g. “You are not correct.”) and elaborated – providing verification and explanation – to help the student understand how to solve the problem. An example task is shown in Figure 4.

To the extent feasible, the elaborated feedback for an incorrect response was “diagnostic” in the sense of being crafted to diagnose misconceptions or procedural bugs suggested by the student’s response. Feedback was written by two of the co-investigators and one other person, all of whom have Ph.D. degrees in cognitive psychology (or equivalent experience) and who have all worked professionally on mathematics projects.

If the learner entered an incorrect answer (e.g. “27”) to the question shown in Figure 4, the elaborated feedback would indicate the following, “That’s not correct. Three times as many emails go out every hour. That means 3 emails go out in the first hour, 9 go out in the second hour, and 27 emails (your response) go out in the third hour. The question asks about the number of emails in the

fourth hour, which would be  $3 \times 3 \times 3 \times 3 = 81$ .” Thus with elaborated feedback, the accuracy of the answer is indicated, along with an explanation about how to solve the problem and the correct answer.

Hour: 1 2 ...

*Emily receives an email message which states that she'll have a "very lucky day" if she sends it out within one hour to exactly 3 people who, in turn, send it out to exactly 3 people, and so on. Emily forwards the email and everyone she sends it to participates in the chain mail. How many emails would be sent at the 4<sup>th</sup> hour? Enter your answer here: \_\_\_\_\_.*

Fig.4. Example ACED geometric sequence item with constructed response required.

## STUDY DESIGN AND METHOD

### Sample

A total of 268 Algebra I students participated in the study. These students attended the same mid-Atlantic state suburban high school. According to their teachers, for these students, geometric sequences were not explicitly instructed as part of the curriculum, although some geometric sequence problems may have been covered as part of other topics in algebra. Testing was conducted in sessions of about 20 students each, over a period of five days. Our full sample of students<sup>7</sup> was heterogeneous in ability, representing a full range of levels of math skills: honors ( $n = 38$ ), academic ( $n = 165$ ), regular ( $n = 27$ ), remedial ( $n = 30$ )<sup>8</sup>, and special education students ( $n = 8$ ).

### Procedure

Each two-hour session consisted of the following activities. First, all students, at their desks, received a 10-minute introduction which included the goals of our study and a description of upcoming activities. We also told them that their participation would not affect their math grade, and that it was

<sup>7</sup> Originally, we tested 290 students, which included a group of 22 English language learners (ELL) with very limited English competency. However, during the posttest, the bilingual teacher of the 22 ELL students assisted them not only in translating the problems, but in many cases, in solving the problems. The teacher was not present during pretesting of the students. Consequently, the ELL students' data were removed from the database as their "learning gains" were artificially inflated. The final sample size for the study is thus 268 students.

<sup>8</sup> Remedial students generally require four years to complete a sequence of courses that others typically require 2-3 years to complete.

important that they try their hardest. Finally, we reminded them that they were getting a reward for their participation.

After the introduction, all students took a 20-minute pretest. We created two test forms for the pre- and posttests – Forms A and B. The tests contained 25 items spanning the identified content (Geometric Sequences) and were administered in paper and pencil format. Items in each form were matched in relation to a specific competency, difficulty level, and format. Calculators were permitted. After the pretest, students were randomly assigned to one of four conditions (see Experimental Conditions, below) where they spent the next one hour either at the computer in one of the three variants of ACED, or at their desks for the control condition. Following the one hour period, all students returned to their desks to complete the 20-minute posttest and a 10-minute paper and pencil survey.

### Design and Experimental Conditions

For the evaluation of ACED, we used a pretest-treatment-posttest design with participating students being randomly assigned to one of four conditions. All individuals regardless of condition received two forms of a multiple choice test – one form as a pretest and the other form as a posttest. The order of forms was randomly assigned so that half the students received forms in A-B order and the other half in B-A order. The control condition involved no treatment but only the pretest and posttest with an intervening one-hour period (the duration of the ACED intervention) sitting at their desks reading content that was not related to math (e.g. other school work and magazines we obtained from the school library). Students assigned to the experimental conditions (Conditions 1, 2, and 3, see Table 1) took their assigned seats at one of the 26 networked computers in the laboratory where all testing occurred. After logging in, they spent the next hour solving geometric sequence problems presented on the screen. The majority (i.e. 80%) of the students completed all 63 tasks, and 95% of all students had 15 or fewer items remaining at the end of one hour.

For the adaptive conditions (1 and 2), tasks presented to the student varied in relation to the student's particular solution history. This was illustrated earlier in Figure 3. For students in the linear-task condition (Condition 3), the order of items was (a) alphabetical by concept (node) name, and (b) within a node, the items increased in difficulty. For example, each node has an associated set of at least six items with three levels of difficulty (1 = easy, 2 = medium, and 3 = hard) and two similar variants (i.e. item 'a' and 'b' denoting task isomorphs). Thus the first six items in the linear sequence were: Common Ratio 1a, Common Ratio 1b, Common Ratio 2a, Common Ratio 2b, Common Ratio 3a, and Common Ratio 3b. See the Appendix for a list of all competencies in order, and examples of items per node.

For *Research Question 1* ("Is elaborated feedback more effective for student learning than simple feedback?"), the main contrast of interest is between Conditions 1 and 2 (holding task sequencing constant while varying type of feedback). Our hypothesis was that the elaborated feedback group (Condition 1) would experience greater learning than the simple feedback group (Condition 2), given the provision of information related to the topic in question. For *Research Question 2* ("Does adaptive sequencing of tasks have a negative impact on learning?"), the main contrast of interest is that between Conditions 1 and 3 (holding feedback type constant while varying task sequencing). Our hypothesis was that the adaptive sequencing of tasks for information gain (Condition 1) would have no negative effect on learning compared to the linear sequencing of tasks (Condition 3). We did not include the

other possible condition (S/L – simple verification feedback with linear sequencing) because our contrasts of interest could be tested with the other three conditions.

Table 1  
Four experimental conditions

Condition	Code	Feedback for Correct	Feedback for Incorrect	Task Sequencing
1. Elaborated feedback/adaptive sequencing	E/A	Verification	Verification + Explanation	Adaptive
2. Simple feedback/adaptive sequencing	S/A	Verification	Verification	Adaptive
3. Elaborated feedback/ linear sequencing	E/L	Verification	Verification + Explanation	Linear
4. Control – no assessment, no instruction	CONTROL	N/A	N/A	N/A

While not used in a key research question, Condition 4, our Control group, serves the useful function of establishing a base level of transfer from pretest to posttest. This condition thus provides a check on the overall impact of any of the three other conditions (1, 2, and 3). The sample sizes for the four conditions were as follows: Condition 1 (E/A:  $n = 71$ ), Condition 2 (S/A:  $n = 75$ ), Condition 3 (E/L:  $n = 67$ ), and Condition 4 (Control:  $n = 55$ )<sup>9</sup>.

## RESULTS

### Learning from ACED

Because our two test forms (A and B) showed slight differences in difficulty (albeit, not significantly so), we scaled the scores for the pretest by producing z-scores for each pretest form (A and B) and then created a basic scale where 50 was the mean and 10 was the standard deviation. This served to equate the two forms of the pretest. Posttest scores (Forms A and B) were then placed on the same scale that was developed for the pretest via conversion tables we created from raw to scale scores for Form A and Form B (i.e. the raw score to z-score transformation was established on the basis of the pretest results only). As the Form A/B assignment was random, this implicitly produces random equivalent groups with equal mean and variance, equating between the two forms. Next, we examined the general question: Do students working with ACED (all 3 conditions, combined) show evidence of learning compared to the control group? An ANOVA was computed using combined ACED data (all three experimental conditions) in relation to the control group. First, there were no differences between the ACED and control groups in terms of their *pretest* scores (pretest:  $F_{1, 266} = 0.23$ ; NS).

<sup>9</sup> Note that the sample size for the control group was slightly less than the other groups for two reasons. First, our randomization procedure involved assigning students a number from 1-4 (starting at 1 in each class) which mapped to the four conditions. Because classes were not equal in size, we had fewer “4s” (control) than other numbers. Second, because there was a disproportionate number of ELL students assigned to the control condition, when that group of students was removed from the study, it reduced the sample size of the control condition more than the other conditions.

However, the *posttest* scores of ACED students ( $M = 56.5$ ,  $SD = 10.7$ ,  $n = 213$ ) were significantly higher than the *posttest* scores of the Control group ( $M = 52.4$ ,  $SD = 11.6$ ,  $n = 55$ ); *posttest*:  $F_{1, 266} = 6.00$ ;  $p < 0.02$ . The effect size was 0.38. (Cohen's  $d$ ). Thus ACED, in general, does appear to increase student learning across the one-hour assessment period.

### Feedback and Adaptivity Effects on Learning

Research questions 1 and 2 address the relationship to learning of (a) elaborated versus simple feedback, and (b) adaptive versus linear sequencing of tasks. Because our sample was so varied in ability level, we included two independent variables in the analysis: condition and academic level. An ANCOVA was computed with *posttest* score as the dependent variable, *pretest* score as the covariate, and Condition (1-4) and Academic Level (1-5)<sup>10</sup> as the independent variables. The main effects of both Condition ( $F_{3, 247} = 3.41$ ,  $p < 0.02$ ) and Level ( $F_{4, 247} = 11.28$ ,  $p < 0.01$ ) were significant, but their interaction was not ( $F_{12, 247} = 0.97$ , NS). Figure 5 shows the main effect of condition in relation to *posttest* (collapsed across academic level) where the best *posttest* performance is demonstrated by students in the E/A condition, which also shows the largest *pretest*-to-*posttest* improvement. Confidence intervals (95%) for the *posttest* data, per condition, are also depicted in the figure.

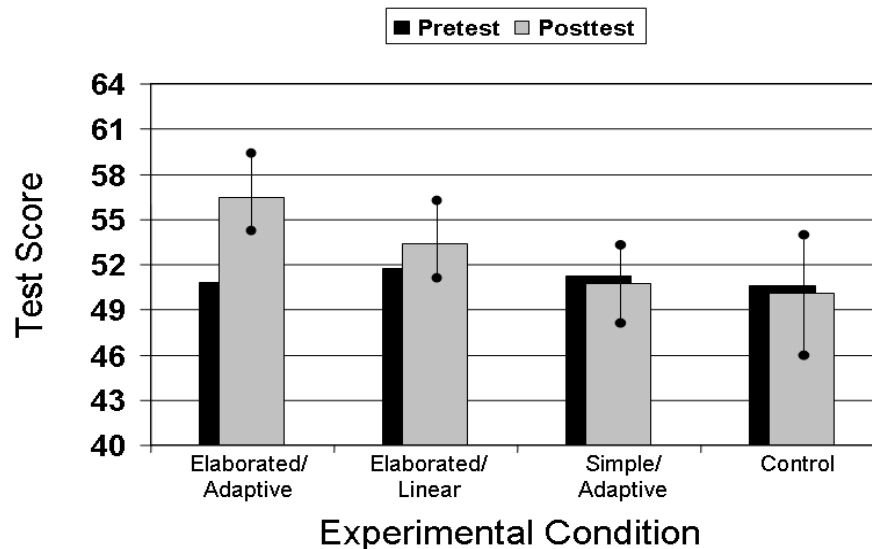


Fig.5. Experimental condition by pre- and *posttest* scores.

The general finding of a main effect of condition on learning prompted a more specific planned comparison (Least Significant Difference, or LSD test) involving all four conditions in relation to *posttest* data. Findings showed that there was a significant difference between the E/A and S/A conditions (i.e. task sequencing the same but type of feedback different). The Mean Difference = 5.74;  $SE = 2.09$ ; and  $p < 0.01$ . There was also a significant difference between the E/A and the Control conditions (Mean Difference = 6.36;  $SE = 2.52$ ; and  $p < 0.02$ ). None of the other comparisons were statistically different. Finally, because the difference between the E/A and the E/L conditions (i.e. type

<sup>10</sup> The five levels included: 1 (honors), 2 (academic), 3 (regular), 4 (remedial), and 5 (special education).



of feedback the same, but task sequencing different) was not significant, this suggests that the elaborated feedback feature was primarily responsible for the impact on learning.

So, regarding the learning questions: (1) The ACED system *does* help students learn – due mainly to the effects of elaborated feedback, and (2) the sequencing of tasks (adaptive vs. linear) did not show significant effects on learning. However, as shown in Figure 5, the E/A vs. E/L conditions suggest a possible advantage for adaptivity. That is, the two conditions were identical except for the sequencing of tasks, and the students assigned to the adaptive version of ACED (E/A) showed higher (although not significantly so) posttest scores than their counterparts in the linear condition (E/L). Also recall that the adaptive algorithm used within ACED was intended to *increase the accuracy of the assessment*, not enhance learning outcome. In the next section we examine assessment-related questions, and present findings showing some benefits of adaptivity.

To be worthwhile, an assessment needs to be reliable and valid. Reliability estimates the *consistency* of an assessment – the degree to which it rank orders students in the same way each time it is used. Validity refers to the extent to which the assessment *accurately* measures what it is supposed to measure. We start by examining the validity of the ACED measurement instruments. This is followed by analyses of reliability of our measures. Finally, we conclude this section with an analysis of efficiency of measurement. That is, while assessments should be both reliable and valid, to be useful they should also be efficient.

## Assessment Quality in ACED

### Validity

The first issue concerns whether the ACED competency estimates (i.e. the values computed by the system and contained within the Bayes net) predict outcome performance beyond that predicted by pretest scores. Each competency variable (see nodes in Figure 1) possesses a triplet of probabilities, reflecting the current estimate of being high, medium, and low on that competency. To reduce the three numbers to a single number, we assigned numeric values +1, 0 and -1 to the three competency states, and computed the expected value. This Expected A Posteriori (EAP) value can also be written as,  $P(\vartheta_{ij} = \text{High}) - P(\vartheta_{ij} = \text{Low})$ , where  $\vartheta_{ij}$  is the value for Student  $i$  on Competency  $j$ , and  $1*P(\text{High}) + 0*P(\text{Med}) + -1*P(\text{Low}) = P(\text{High}) - P(\text{Low})$ . This results in a scale ranging from -1 to 1. As shown in Figure 1, all lower-level competency nodes feed evidence into the main node – Solve Geometric Sequences (SGS). Thus, EAP(SGS) is the overall “score” from ACED, and higher values of EAP(SGS) should be associated with greater knowledge and skills overall on geometric sequence topics (i.e. better performance on the posttest).

A regression analysis was computed with posttest score as the dependent variable and (a) pretest score and (b) EAP(SGS) as the independent variables.<sup>11</sup> Pretest score was forced into the equation first, followed by EAP(SGS). This regression analysis was intended to provide a general sense of whether the ACED estimates were valid or not, and whether they accounted for any unique outcome variance beyond that attributable to pretest scores. Results showed that both independent variables

---

<sup>11</sup> Simple correlations among the three main variables – pretest, posttest, and EAP(SGS) – show that: (1) pretest x posttest,  $r = .59$ , (2) pretest x EAP(SGS),  $r = .50$ , and (3) posttest x EAP(SGS),  $r = .65$ , all significant at  $p < 0.01$ .

significantly predicted 50% of the posttest variance (Multiple  $R = .71$ ;  $F_{2, 210} = 106.57$ ;  $p < 0.001$ ). More specifically, the pretest score predicted 33% of the outcome variance (Multiple  $R = .58$ ;  $p < 0.001$ ), and EAP(SGS) accounted for an *additional* 17% of the outcome variance (Multiple  $R = .42$ ;  $p < 0.001$ ). To test whether this additional (17%) outcome variance is significantly more than that predicted by the pretest (33%), the  $R^2$  difference test refers to running regression for a full model and for the model minus one variable, and then subtracting the  $R^2$  values and testing the significance of the difference. The  $R^2$  increments are tested by the F-test. In this case, the increase of 0.17 over 0.33 was, in fact, significant ( $F_{1, 210} = 73.07$ ;  $p < 0.001$ ).

### ***Reliability of ACED Tasks***

All 63 tasks in the ACED pool of geometric items were statistically linked to relevant competencies. Students in all three ACED conditions were required to spend one hour on the computer solving the 63 items. This design decision was made in order to control for time and let outcomes vary, thus students in the two adaptive conditions spent the same amount of time on the program as those in the linear condition. Because students in all conditions had to complete the full set of 63 items, we obtained accuracy data (scored as 0/1 for incorrect/correct) per student, per item. These performance data were analyzed using a split-half reliability procedure (via SPSS). The Spearman Brown split-half reliability with unequal halves (i.e. 31 and 32) was equal to 0.84, while Cronbach's  $\alpha = 0.88$  which is quite good.<sup>12</sup>

### ***Reliability of ACED Competency Estimates***

We analyzed competency estimates from the Bayes net, again using task performance data which provided input to posterior probabilities, per node. These probabilities were then analyzed, making use of split-half reliabilities at the node level.<sup>13</sup> The reliability for the parent competency – EAP(SGS) – was 0.88 (see Table 2). Moreover, the other competency estimates showed equally impressive reliabilities for their associated tasks. The same pattern held with data using probabilities – triplets per node – and MAPs (Maximum A Posteriori values). The reliabilities obtained for the competency estimates show how we can borrow strength and augment reliabilities in a meaningful way from indirect sources in the competency model even though the “tests” are very small (i.e. the number of tasks per node is between 6-10 tasks). Furthermore, the high reliabilities of the lower-level competencies suggest that they may be employed for diagnostic purposes.

---

<sup>12</sup> A widely-used rule of thumb of 0.70 has been suggested as a minimally-accepted internal reliability score (e.g. Nunnally, 1978), and an ideal estimate of internal consistency is between 0.80 and 0.90 because estimates above .90 are suggestive of item redundancy or inordinate scale length (e.g. Clark & Watson, 1995). This heuristic is useful for both the Cronbach's  $\alpha$  and the Spearman-Brown split-half reliability estimates (which are on the same correlation scale).

<sup>13</sup> Here we were able to make use of the ECD design information to make two half-forms that were almost exactly parallel, i.e. tasks were matched on target proficiency and difficulty. All but a few tasks had an isomorph in both halves.

Table 2  
Spearman-Brown split-half reliabilities

Competency (EAP)	Reliability
<i>Solve Geometric Sequences (SGS)</i>	0.88
Find Common Ratio	0.90
Generate Examples	0.92
Extend Sequence	0.86
Model Sequence	0.80
Use Tables	0.82
Use Pictures	0.82
Induce Rules	0.78

### ***Reliability of ACED Pretest and Posttest (Forms A and B)***

Separate analyses were computed on the reliabilities of the pretest (Forms A/B) and posttest (Forms A/B) items. We computed Cronbach's  $\alpha$  for each of the four tests, then used Spearman Brown's prophecy formula to increase the size of the tests to 63 items to render the tests comparable in length to the ACED assessment. These reliabilities are shown in Table 3.

Table 3  
Reliabilities across pretests and posttests

Test Form	Pretest $\alpha$	Posttest $\alpha$
Adjusted Form A	0.84	0.82
Adjusted Form B	0.79	0.87
<b>Adjusted Avg (A/B Combined)</b>	<b>0.82</b>	<b>0.85</b>

### ***Efficiency of the ACED System***

In this study, we required that students in all three ACED conditions spend one hour on the computer solving the set of 63 items. As noted earlier, the majority (i.e. 80%) of the students completed all 63 tasks, and 95% of all students had 15 or fewer items remaining at the end of one hour. Those who completed the program early returned to their seats to read or rest until the hour was up. Even though two of our ACED conditions employed our adaptive algorithm for task selection (i.e. Conditions 1 and 2 – E/A and S/A), we still required students to complete the full set of 63 tasks. A typical rationale for using adaptive tests, however, relates to their efficiency. That is, adaptive algorithms rely on fewer tasks or items to determine competency level than more traditional approaches. The question here concerns what the data (competency estimates) would look like if we required fewer tasks to be solved (i.e. implementing a termination criterion into the adaptive algorithm). To answer this question, we selected the first  $N$  (where  $N = 10, 15, 20, 25, 30, 40, 50,$  and  $63$ ) tasks from the student records, and then calculated EAP values for the parent competency (Solve Geometric Sequences) from each "shortened" test. Next, we computed correlations of each of these tests with the posttest score for the students.

What we expected to see was that the correlations, in general, should increase with test length until it reaches an upper asymptote related to the reliability of the posttest. We hypothesized that the data from students in the linear condition (E/L) should reach that asymptote more slowly than the data

from participants in the adaptive conditions. Figure 6 shows the results of the plot, confirming our hypothesis. The quick rise and asymptote of the two adaptive conditions shows that only 20-30 tasks are needed to reach the maximum correlation with the posttest. At that juncture, for those students, the next step would be instructional intervention by the teacher or computer program.

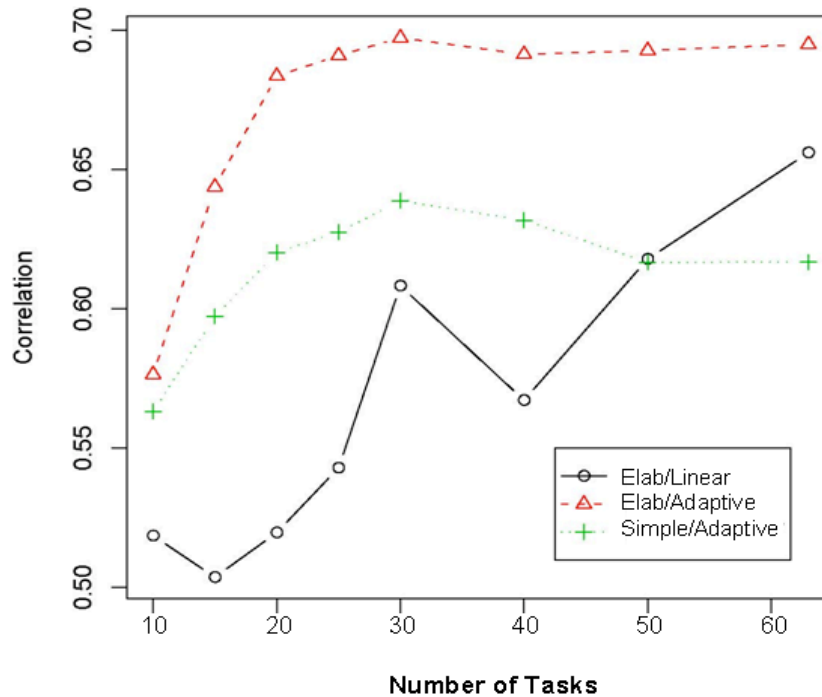


Fig.6. Correlations of EAP (SGS) with posttest score by ACED condition.

In the middle of the linear condition (E/L), there is a spike around task 30, then a subsequent drop, interrupting the gradual climb of the E/L curve. This is likely caused by a specific pattern of items, e.g. a collection of easy items which were good predictors followed by a collection of hard items which were poor predictors. In fact, when we reviewed the list of the 63 items in the order in which they appeared in the linear condition, items 31-36 consisted of a group of rather difficult items, “Determine the recursive rule for geometric sequences”. An example of a task linked to that particular competency (i.e. *recursive rule*, a low-level node shown in Figure 1) is:

Let  $a_1, a_2, a_3, \dots, a_n, \dots$  be a sequence of integers. The first four terms of the sequence are:

**-6, 24, -96, 384, ...** Select the correct rule for finding the  $n+1^{\text{th}}$  term from the  $n^{\text{th}}$  term.

- (a)  $a_{n+1} = -6a_n$
- (b)  $a_{n+1} = 4a_n$
- (c)  $a_{n+1} = -a_n/6$
- (d)  $a_{n+1} = a_n \cdot 4^{n-1}$
- (e)  $a_{n+1} = -4a_n$

Finally, notice the slight decline after task 30 in the S/A condition. This could represent a fatigue effect. That is, because many of the problems were unfamiliar, and the feedback in this condition only

indicated correct or incorrect (with no help or instruction), students may have stopped trying their best. Thus another potential benefit of the elaborated feedback could be student engagement. However, the noise associated with these data is large (e.g. the 95% confidence interval for the correlation coefficients ranges from .46 to .75) requiring another independent verification of this decline phenomenon.

## SUMMARY AND DISCUSSION

The study described in this paper evaluated ACED as both (a) a valid and reliable assessment instrument, and (b) a learning aid. The program succeeded in both capacities. Despite the fact that ACED does not adjust its competency estimates to reflect student learning during the assessment, it nonetheless produces a valid assessment quite efficiently. Based on the results described herein, we have an answer to our earlier question, “*When students are given feedback on their task solutions, does their learning render the assessment less valid, reliable, or efficient?*” The answer is “No.” Now, a more direct test of this hypothesis would employ a condition where ACED adaptively sequenced the tasks but did not give any feedback (No-feedback/A). However, we would argue that because the E/A condition produced more learning than the S/A condition, and their validity, reliability, and efficiency indices were unaffected by the elaborated vs. simple feedback conditions, this strongly suggests that if *no feedback* were given, the validity, reliability, and efficiency indices would be about the same as the E/A and S/A versions of ACED which *did* give feedback. In addition, the ACED scores showed higher correlation with the posttest scores compared to the pretest-posttest correlation, even though the pretest and posttest were designed to be parallel. Finally, and contrary to expectations, increasing the feedback from simple verification to elaborated feedback actually *increased* the correlation with the posttest (albeit, not significantly), as shown in Figure 6. It is possible that providing feedback helps reduce guessing on the part of students who are at the cusp of mastery for a certain competency. This hypothesis should be investigated with future studies.

Overall, students in all three ACED conditions (combined) demonstrated significantly more learning of the topic of Geometric Sequences than those assigned to the control condition. More importantly, we identified the source of the learning effects – elaborated feedback. That is, we found a significant difference in posttest scores between students in the E/A and S/A conditions. Both groups received adaptively-sequenced tasks, but they differed in the type of feedback received, with a clear benefit of elaborated feedback. There was no significant main effect of adaptivity (E/A vs. E/L), although we reiterate that the intention of the adaptive algorithm was to enhance the precision of the assessment than to facilitate learning. The trend displayed in Figure 5, however, showed that students in the adaptive E/A condition ended up with higher posttest scores than those in linear E/L condition, although this was not a statistically significant difference.

Results regarding predictive validity suggest that the use of evidence-based assessment design with Bayes net technology may be a valuable approach for estimating competencies from performance data. Regression analysis showed just a single competency estimate can significantly predict posttest performance, beyond that provided by the pretest. Next, the split-half reliability of the 63 ACED tasks was high (0.88), as was the parent competency (EAP (SGS)) reliability (0.88) and our adjusted pretest and posttest reliabilities (0.82 and 0.85, respectively). The results also suggest that for students in the adaptive conditions (i.e. E/A and S/A), the ACED program could have terminated after approximately 20 items without degradation in prediction of outcome. This hypothesis would require additional study

to validate (e.g. conducting a follow-up study that stops the program after 20 items for the adaptive ACED conditions, and administers an intermediate test). Finally, results of the study tend to confirm expectations regarding the efficiency of adaptive assessment relative to linear assessment.

Regarding next steps, we would like to enhance our adaptive algorithm to better support student learning. The current adaptive algorithm was designed mainly to enhance assessment, but ACED – which incorporates elaborated feedback – was also intended to enhance learning. That is, while the expected weight of evidence algorithm has certain optimality properties for assessing student competencies, it does not necessarily have optimal properties for supporting *growth* in those competencies, i.e. student learning. The algorithm tends to select tasks which the student has an approximately 50% chance of getting right, but the idea that such tasks provide good learning opportunities needs more investigation.

We would also like to improve our modeling of student learning. One limitation of the current approach is that the ACED scoring engine assumes that the student's competency does not change over the course of the ACED session, but that does not seem to be true. To illustrate the problem, suppose a student is struggling with one of the competencies represented in Figure 1 (e.g. generate examples of geometric sequences). The student receives one medium and two easy items linked to this competency and solves all three incorrectly. Later, something from the elaborated feedback provokes an "Aha" moment and the student subsequently solves two easy items and one medium item correctly. However, ACED treats this as a cumulative set of six items: 3 right, 3 wrong, ignoring the temporal sequence, and assuming that the student's competency is the same throughout the assessment period. The resulting probability distribution for this particular competency would likely be flat, and the EAP estimate would dip below zero and then come back up to zero, but rather slowly. But if ACED considered the temporal ordering of the data, the earlier observations would be discounted in some way, and the current EAP estimate would be positive. A more sophisticated engine needs to take student growth into account in a realistic way. Ideally we should be able to use the inferences from the Bayes net as input to a planning system to help suggest next steps for the teacher, computer system, or student. See Almond (2007a) for a more detailed discussion of this topic.

One final and important question concerns the scalability of this type of evidence-based assessment for learning approach. In general, assessments that focus on a particular topic tend to have higher reliability than those focusing across topics. That was the case with this current study that assessed the specific topic of Geometric Sequences. However, even if broadening the span of ACED would slightly decrease its reliability, the methodology should scale easily to larger competency models and collections of tasks. In general, the computational cost of Bayesian networks is linear in terms of the number of nodes, but exponential in the number of edges coming into each node. Thus, the limits of the system are less strongly influenced by the number of competency variables or the number of tasks (although this would slightly increase the searching cost for the weight of evidence algorithm), but would be influenced by the number of competencies required to solve each task. So, ACED could easily scale to larger domains, but in such cases it is likely that students would be unable to complete the assessment in one sitting. In this case, models that account for student competency growth between sessions become increasingly important.

In conclusion, the policy implications of the findings reported in this paper are potentially quite significant. For instance, our results suggest that many of those countless hours of state-mandated testing might be turned into real learning experiences without adversely affecting the tests' capability to accurately assess the students. As we have pointed out, this represents a persistent dream of some educators (including the authors), dating back to at least Snow and Mandinach (1991). To our

knowledge, this is the first research paper to produce concrete evidence that the dream can become a reality. In general, we envision a role for the ACED program as part of a larger instructional support system. The adaptive assessment for learning system would continue to accurately assess the student until the best instructional options for that student become clear. Meanwhile, elaborated feedback would ensure that the assessment itself was an effective learning experience, supporting learning for students with a wide range of abilities. In short – it seems that we can, in fact, “fatten the hog” with the same instrument used to weigh it.

## ACKNOWLEDGEMENTS

We gratefully appreciate the support for ACED development and data collection by National Science Foundation Grant No. 0313202, especially John Cherniavsky, our program officer. We also want to acknowledge the various contributions to the development of ACED by the following: Edith Aurora Graf, Jody Underwood, Peggy Redman, Malcolm Bauer, Buz Hunt, Diego Zapata-Rivera, Irvin Katz, and Dan Eignor. In addition, we are grateful for the contributions from the New Jersey middle- and high school mathematics teachers who assisted us with content, and to Angelique Peterson (head mathematics teacher at the high school where we tested ACED). She organized the Algebra I students for us to test, and Waverely Hester provided excellent support during the testing as well as during preliminary data analyses. Finally, we appreciate the thorough and thoughtful comments and suggestions made by Kurt VanLehn and two anonymous reviewers of this paper.

## REFERENCES

- Almond, R. G. (1995). *Graphical Belief Modeling*. Boca Raton, FL: CRC Press, Inc.
- Almond, R. G. (2007a). *An illustration of the use of Markov decision processes to represent student growth (learning)*. ETS Research Report, RR-07-40 (pp. 1-61), Princeton, NJ.
- Almond, R. G. (2007b). I can name that Bayesian network in two matrixes! *CEUR Workshop Proceedings*, 268, Retrieved January 7, 2008, from <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-268/paper1.pdf>
- Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, 23, 223-238.
- Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J-D. (2007). Modeling diagnostic assessment with Bayesian networks. *Journal of Educational Measurement*, 44, 341-359.
- Almond, R. G., Yan, D., Matukhin, A., & Chang, D. (2006). *StatShop testing*. ETS, Research Memorandum 06-04, Princeton, NJ.
- Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research*, 13(2), 111-127.
- 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.
- Barr, A., & Feigenbaum, E. (1981). *The Handbook of Artificial Intelligence, Volume 1*. Los Altos, CA: William Kaufmann, Inc.
- Bass, K. M., & Glaser, R. (2004). Developing assessments to inform teaching and learning (CSE Rep. No. 628). Los Angeles, CA: UCLA, Graduate School of Education and Information Studies. Retrieved January 7, 2008, from [http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content\\_storage\\_01/0000019b/80/1b/9d/fb.pdf](http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/9d/fb.pdf).

- Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., & Wiesemes, R. (2006 March/April). A learning integrated assessment system. In R. Wiesemes & G. Nickmans (Eds.) *EARLI Series of Position Papers*.
- Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-71.
- Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-148.
- Brookhart, S. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education: Principles, Policy & Practice*, 8(2), 153-169.
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7, 309-319.
- Clarke, S. (2001) *Unlocking formative assessment: Practical strategies for enhancing pupils' learning in the primary classroom*. London: Hodder & Stoughton.
- Conati, C., Gertner, A. S., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, 12(4), 371-417.
- Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-de-la-Cruz, J. L., & Rios, A. (2004). SIETTE: A Web-Based Tool for Adaptive Teaching. *International Journal of Artificial Intelligence in Education*, 14, 29-61.
- Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. *Proceedings of ACM CHI 2001 Conference on Human Factors in Computing Systems* (pp. 245-252). New York: ACM Press.
- Good, I. J. (1985): Weight of evidence: A brief survey. In J. Bernardo, M. DeGroot, D. Lindley & A. Smith (Eds.) *Bayesian Statistics 2* (pp. 249-269). Amsterdam: North Holland.
- Good, I. J., & Card, W. (1971). The diagnostic process with special reference to errors. *Method of Inferential Medicine*, 10, 176-188.
- Hansen, E. G., & Mislevy, R. J. (2005). Accessibility of computer-based testing for individuals with disabilities and English language learners within a validity framework. In M. Hricko & S. Howell (Eds.) *Online assessment and measurement: Foundation, challenges, and issues*. Hershey, PA: Idea Group Publishing, Inc.
- Hansen, E. G., & Shute, V. J. (2007, April). *Towards accessible educational products: The usability of an assessment-for-learning system in mathematics for individuals with visual disabilities*. Paper presented at the annual National Council on Measurement in Education meeting, Chicago, IL, April, 2007.
- Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests for individuals with disabilities within a validity framework. *System: An International Journal of Educational Technology and Applied Linguistics*, 33(1), 107-133.
- Harlen, W. (2005). Teachers' summative practices and assessment for learning – tensions and synergies. *Curriculum Journal*, 16(2), 207-223.
- Hensen, R., & Douglas, J. (2005) Test construction for cognitive diagnosis. *Applied Psychological Measurement*, 29, 262-277.
- Jameson, A. (2007). Adaptive interfaces and agents. In J. A. Jacko & A. Sears (Eds.) *The human-computer interaction handbook: Fundamentals, evolving technologies & emerging applications* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254-284.
- Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: Minute by minute, day by day. *Educational Leadership*, 63(3), 19-24.
- Lepper, M. R., & Chabay, R. W. (1985). Intrinsic motivation and instruction: Conflicting views on the role of motivational processes in computer-based education. *Educational Psychologist*, 20(4), 217-230.
- Madigan, D., & Almond, R. G. (1996). On test selection strategies for belief networks. In D. Fisher & H.-J. Lenz (Eds.) *Learning from data: AI and statistics IV* (pp. 89-98). New York: Springer-Verlag.

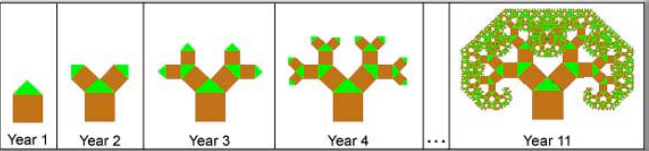


- Madigan, D., Mosurski, K., & Almond, R. G. (1997). Graphical explanation in belief networks. *Journal of Computational and Graphical Statistics*, 6(2), 160-181.
- Mason, B. J., & Bruning, R. (2001). *Providing feedback in computer-based instruction: What the research tells us*. Retrieved January 2, 2008, from <http://dwb4.unl.edu/dwb/Research/MB/MasonBruning.html>.
- McTighe, J., & O'Connor, K. (2005). Seven practices for effective learning. *Educational Leadership*, 63(3), 10-17.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher*, 23(2), 13-23.
- Millán, E., & Pérez-de-la-Cruz, J. L. (2002). A Bayesian diagnostic algorithm for student modeling and its evaluation. *User Modeling and User-Adaptive Interaction*, 12, 281-330.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the role of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.) *Item generation for test development* (pp. 97-128). Mahwah, New Jersey: Lawrence Erlbaum.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-62.
- Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science*, 32, 99-113.
- Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multi-media learning. In H. M. Niegemann, D. Leutner & R. Brunken (Eds.) *Instructional design for multimedia learning* (pp. 181-195). Munster, New York: Waxmann.
- Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70(350), 351-371.
- Peat, M., & Franklin, S. (2002). Supporting student learning: the use of computer-based formative assessment modules. *British Journal of Educational Technology*, 33(5), 515-523.
- Popham, W. J. (2001). Teaching to the test. *Educational Leadership*, 58(6), 16-20.
- Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., et al. (2005). The Assistent Project: Blending assessment and assisting. In C. K. Looi, G. McCalla, B. Bredeweg & J. Breuker (Eds.) *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology* (pp. 555-562). Amsterdam: IOS Press.
- Schwarz, R. D., & Sykes, R. (2004, June). *Psychometric foundations for formative assessment*. Paper presented at the National Educational Computing Conference, New Orleans, LA.
- Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press.
- Shepard, L. A. (2003). Reconsidering large-scale assessment to heighten its relevance to learning. In J. M. Atkin & J. E. Coffey (Eds.) *Everyday assessment in the science classroom*. Arlington, VA: NSTA Press.
- 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: 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., & Zapata-Rivera, D. (in press). Educational measurement and intelligent systems. In E. Baker, B. McGaw & P. Peterson (Eds.) *Third Edition of the International Encyclopedia of Education*. Oxford, UK: Elsevier Publishers.
- 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 Edition) (pp. 277-294). New York, NY: Lawrence Erlbaum Associates, Taylor & Francis Group.
- 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. (2007). *An assessment for learning system called ACED: Designing for learning effectiveness and accessibility*. ETS Research Report, RR-07-26 (pp. 1-54), Princeton, NJ.
- Snow, R. E., & Mandinach, E. B. (1991). *Integrating assessment and instruction: A research and development agenda*. Princeton, NJ: Educational Testing Services.
- Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan Professional Journal*, 83(10), 758-765.
- Stiggins, R. (2006). Assessment for learning: A key to motivation and achievement. *Edge: The Latest Information for the Education Practitioner*, 2(2), 1-19.
- Symonds, K. W. (2004). *After the test: Closing the achievement gaps with data*. Naperville, IL: Learning Point Associates. Retrieved January 3, 2008, from <http://www.ncrel.org/gap/studies/basrc.pdf>
- Taras, M. (2002). Using assessment for learning and learning from assessment. *Assessment and Evaluation in Higher Education*, 27(6), 501-510.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.
- Tatsuoka, K. K. (1985). A probabilistic model for diagnostic misconceptions in the pattern classification approach. *Journal of Educational Statistics*, 10, 55-73.
- Tiknaz, Y., & Sutton, A. (2006). Exploring the role of assessment tasks to promote formative assessment in key Stage 3 Geography: Evidence from twelve teachers. *Assessment in Education: Principles, Policy and Practice*, 13(3), 327-343.
- Van der Linden, W. J. (1998) Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2), 201-216.
- VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3), 227-265.
- Vygotsky, L. (1967). Play and its role in the mental development of the child. *Soviet Psychology*, 5(3), 6-18.
- Wainer, H. (1990). *Computer adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2<sup>nd</sup> Edition). Mahwah, NJ: Erlbaum Associates.
- Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.
- Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: impact on student achievement. *Assessment in Education: Principles, Policy & Practice*, 11(1), 49-65.
- 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* (Vol. 158, pp. 323-330). Amsterdam: IOS Press.

**APPENDIX: Competencies in ACED (Geometric Sequences)**

<b>Competency</b>	<b>Definition</b>	<b>Example Task (Difficulty = Medium)</b>										
<i>Common Ratio</i>	Find the common ratio in a geometric sequence of numbers.	Find the common ratio for the following geometric sequence: 5, -10, 20, -40, 80, . . . (Answer: -2)										
<i>Examples</i>	Generate terms of a geometric sequence that satisfy the constraints provided by giving the first few terms of the geometric sequence.	Give the first 3 terms of a geometric sequence that satisfies the following: <ol style="list-style-type: none"> <li>The first term is a negative integer between -6 and -4, inclusive.</li> <li>The common ratio is a positive integer between 4 and 6, inclusive.</li> </ol> (Answer: There is more than one correct answer. Any example you choose that meets the above requirements will be scored as correct.)										
<i>Explicit Rule</i>	Generate an algebraic expression that represents the $n$ th term for the given geometric sequence.	The first four terms of a geometric sequence are given below. Choose the algebraic expression that represents the $n$ th term in this sequence. 27, 81, 243, 729, . . . <ol style="list-style-type: none"> <li><math>3n + 27</math></li> <li><math>5 \times 3^n</math></li> <li><math>3^{n+2}</math></li> <li><math>3n</math></li> <li><math>27^n</math></li> </ol> (Answer: $3^{n+2}$ )										
<i>Extend</i>	Extend the geometric sequence that has simple starting terms and common ratio.	Find the missing terms in the following geometric sequence: <b>2, __, 18, __, 162,</b> (Answer: 6, 54)										
<i>Identify Geometric</i>	Identify which of the following sequences is not a geometric sequence.	Which of the following is not a geometric sequence? <ol style="list-style-type: none"> <li>6, 5, 5, 4, . . .</li> <li>2, 1, __, __, . . .</li> <li>1, 3, 9, 27, . . .</li> <li>3, 6, 12, 24, . . .</li> </ol> (Answer: 6, 5, 5, 4, . . .)										
<i>Model</i>	Generate a geometric sequence that represents the given situation.	At a certain bakery chain, fresh baked goods are prepared every morning. The number of pounds of baked goods that the bakery chain sells per hour, starting with the hour the bakeries open, is shown in the chart below. Pounds of Baked Goods sold per hour <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Hour</th> <th>Pounds</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>384</td> </tr> <tr> <td>2</td> <td>192</td> </tr> <tr> <td>3</td> <td>96</td> </tr> <tr> <td>4</td> <td>48</td> </tr> </tbody> </table> If this pattern of sales continues, during which hour will the bakery chain first sell less than 15 pounds of baked goods? (Answer: 6)	Hour	Pounds	1	384	2	192	3	96	4	48
Hour	Pounds											
1	384											
2	192											
3	96											
4	48											

<i>Pictorial</i>	Use geometric sequences to model or extend visual patterns provided.	<p>Nicole receives a mysterious looking tree as a housewarming gift. She plants the tree in her garden. Each year, the tree grows new shoots. The pictures below show the pattern of growth across the years.</p>  <p>Observe the first four years of growth. Assuming the same pattern continues, how many <b>new</b> shoots appear in year 11? (Answer: 1024)</p>										
<i>Recursive Rule</i>	Generate or recognize a recursive formula for a given geometric sequence.	<p>Let <math>a_1, a_2, a_3, \dots, a_n</math>, be a sequence of integers. The first four terms of the given sequence are: -3, -6, -12, -24. Select the correct rule for finding the <math>(n+1)^{\text{th}}</math> term from the <math>n^{\text{th}}</math> term.</p> <p>a. <math>a_{n+1} = -2a_n</math>  b. <math>a_{n+1} = 2a_n</math>  c. <math>a_{n+1} = a_n \times 2^{n-1}</math>  d. <math>a_{n+1} = a_n + 2^{n-1}</math>  e. <math>a_{n+1} = (-a_n)/3</math></p> <p>(Answer: <math>a_{n+1} = 2a_n</math>)</p>										
<i>Table</i>	Enter missing values in a table representing a geometric sequence.	<p>The numbers in the table represent terms in a geometric sequence. Complete the table by filling in the values for <b>A</b> &amp; <b>B</b>.</p> <table border="1" data-bbox="730 1043 948 1200"> <thead> <tr> <th>Term Number</th> <th>Term Value</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>1/2</td> </tr> <tr> <td>2</td> <td>1/8</td> </tr> <tr> <td>3</td> <td><b>A</b></td> </tr> <tr> <td>4</td> <td><b>B</b></td> </tr> </tbody> </table> <p>(Answer: <math>A = 1/32, B = 1/128</math>)</p>	Term Number	Term Value	1	1/2	2	1/8	3	<b>A</b>	4	<b>B</b>
Term Number	Term Value											
1	1/2											
2	1/8											
3	<b>A</b>											
4	<b>B</b>											
<i>Verbal Rule</i>	Generate a verbal rule for a geometric sequence.	<p>You have an excellent recipe for chocolate chip cookies. By the end of a week, you've shared the recipe with four of your friends. During the second week, each of your friends shared the recipe with four of their friends, so that sixteen new people know about the recipe. Assuming this pattern continues, how many new friends will get the recipe during the 10th week?</p> <p>a. <math>4^{10}</math>  b. <math>4^9</math>  c. <math>4^8</math>  d. 1,398,101</p> <p>(Answer: <math>4^{10}</math>, which equals 1,048,576.)</p>										