Evaluating ACED: The Impact of Feedback and Adaptivity on Learning

Valerie J. SHUTE[1], Eric G. HANSEN and Russell G. ALMOND
Educational Testing Service, Princeton, New Jersey, USA

Abstract. This paper reports on the evaluation of a program named ACED (Adaptive Content with Evidence-based Diagnosis)—an assessment for learning system using Algebra I content related to the topic of geometric sequences. We used an evidence-centered design (ECD) approach [1] to create the system which includes three main models: proficiency, evidence, and task. Our goals of this study were to determine the learning benefit of the key design elements of adaptivity and formative feedback. Results from an experiment testing 268 heterogeneous students generally show that the system: (a) significantly improves learning, and (b) is a reliable and valid assessment tool. More specifically, the system's formative feedback feature significantly enhanced student learning, but did not detract from the accuracy of the assessment.

Keywords: Assessment for learning, Bayesian networks, evidence-centered design

1. Introduction

An assessment *for* learning (AfL) approach to education involves weaving assessments directly into the fabric of the classroom and using results as the basis to adjust instruction to promote student learning in a timely manner. This type of assessment contrasts with the more traditional, summative approach (i.e., assessment of learning), which is administered less frequently than AfL and is usually used for accountability purposes. In the past decade or so, AfL has shown great potential for harnessing the power of assessments to support learning in different content areas [2] and for diverse audiences. Unfortunately, while assessment of learning is currently well entrenched in most educational systems, assessment for learning is not.

In addition to providing teachers with evidence about how their students are learning so that they can revise instruction appropriately, AfL systems may directly involve students in the learning process, such as by providing feedback that will help students gain insight about how to improve [3], and by suggesting (or implementing) instructional adjustments based on assessment results [4]. These promises, however, need controlled evaluations to determine which features are most effective in supporting learning in a range of settings [5]. Among the AfL features that have the greatest potential for supporting student learning and which would be suitable for investigation are task-level feedback and adaptive sequencing of tasks.

[1] Corresponding Author: Valerie Shute, ETS, Rosedale Rd. Princeton, NJ, USA; Email: vshute@ets.org

1.1. Task-level Feedback

Task-level feedback appears right after a student has finished solving a problem or task, and may be contrasted with (a) general summary feedback which follows the completion of the entire assessment, and (b) specific step-level feedback which may occur within a task [6], like ITSs often provide. Task-level feedback typically gives specific and timely (usually 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., [7, 8, 9]), and may also influence motivation (e.g., [10, 11]). Immediate feedback on students' solutions to individual tasks has generally been shown to support student learning [12, 3], 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 for incorrect answers, research has suggested that it is more beneficial to provide not only verification of the incorrectness but also an explanation of how to determine the correct answer (e.g., [13, 14]). In this research, we focused on feedback for incorrect answers and evaluated the contribution to learning that elaborated feedback provides relative to simple verification feedback.

1.2. Adaptive Sequencing of Tasks

Adaptive sequencing of tasks contrasts with linear (i.e., fixed) sequencing and involves 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 proficiency level (assessment), and (b) which task would be most helpful in supporting the student's progress to a higher proficiency level (learning support). The idea of tailoring content to fit the needs of learners is quite appealing and is being incorporated into many e-learning systems, but lacks clear empirical support [5, 15]. These factors motivated the current research. Specifically, we want to experimentally test the value that adaptive task sequencing adds to learning outcome and efficiency compared to fixed (linear) sequencing of tasks.

1.3. Purpose

This paper describes the evaluation of an AfL system—ACED—that combines feedback and adaptive task sequencing to support students learning of Algebra I content while concurrently assessing their knowledge and skills. Thus the general goal of the evaluation was to obtain answers about whether such an AfL system works—both as an assessment tool and to support learning. Three main features of ACED were tested: feedback type, task sequencing, and proficiency estimation. The primary research questions we examine in this paper include the following: (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)? (2) Is adaptive sequencing of tasks more effective for learning than linear sequencing? (3) Does the ACED system provide reliable and valid information about the learner?

2. Methodology

For this evaluation, 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 mathematics. 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.

For Research Question 1 ("Is elaborated feedback more effective for student learning than simple feedback?"), the main contrast of interest is that 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). For Research Question 2 ("Is adaptive sequencing of tasks more effective for learning than linear sequencing?"), 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 group (Condition 1) would experience greater learning than the linear sequencing group (Condition 3). While not used in a key research question, Condition 4, our Control group, serves the useful function of establishing a base level of learning resulting from no assessment-for-learning. This condition thus provides a check on the overall impact of any of the three other conditions (1, 2, and 3).

Table 1. Four Conditions Used in the Experiment

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

2.1. Content

The topic area of "sequences" (e.g., arithmetic, geometric, and other recursive sequences, like the Fibonacci series) was selected for implementation based on interviews with school teachers, review of state standards in mathematics, and so on. ECD models (proficiency, evidence, and task) were developed for the full set of proficiencies, but the evaluation described in this paper focuses on one branch of the proficiency model—geometric sequences (i.e., successive numbers linked by a common ratio). Shute and colleagues [16, 21] describe ACED more fully, including the proficiency model for geometric sequences, which was expressed as a Bayesian network with 8 main nodes corresponding to the key proficiencies. For each node in the

proficiency model there were at least six tasks available to administer to the student—three levels of difficulty and two parallel tasks per level.

2.2. Adaptive Algorithm

We integrated an adaptive algorithm into the program to determine which item to present next to maximize measurement accuracy. This involved calculating the expected weight of evidence [17, 18] per task as the basis for selecting subsequent tasks. That is, consider any hypothesis we might want to make about a student's ability—e.g., that the student's Solve Geometric Sequences proficiency (the parent node in the model) is at or above the medium level. At any point in time, the Bayes net can calculate the probability 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 will result in a change in the probabilities. The change in log odds is called the weight of evidence for the hypothesis provided by the evidence [19]. Typically, conclusions about a hypothesis are based on the accumulation of many bits of evidence. For outcomes from tasks that have not been observed, one can calculate the expected weight of evidence under the hypothesis [17]. The task that maximizes the expected weight of evidence (i.e., the most informative about the hypothesis) will be chosen. This approach is related to the procedures commonly used in computer-adaptive testing (e.g., [20, 21]), but differs in that ACED explicitly employs feedback to support learning as part of the assessment.

2.3. Authoring Tasks

A set of 63 tasks (questions) were authored for the geometric sequences branch of the proficiency model. Each task was statistically linked to relevant proficiencies. 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 writing feedback for multiple-choice answers and for common errors to constructed response items. Verification feedback was used for correct answers (e.g., "You are correct!"). If the response was incorrect, the feedback provided elaboration—verification and explanation—to help the student understand how to solve the problem. To the extent feasible, the elaborated feedback for an incorrect response was "diagnostic" in the sense of being crafted based on a diagnosis of the misconception or procedural bug suggested by the student's response. If the learner entered an incorrect answer, the feedback in the simple feedback condition (S/A) would note, "Sorry, that's incorrect" and the next item would be presented. In the elaborated feedback condition (E/A), the computer would respond to the incorrect answer with the accuracy of the answer, a short and clear explanation about how to solve the problem, and often the correct answer (see [16] for specific examples of tasks, feedback, and other system details).

2.4. 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 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), and special education students (n = 8).

2.5. Procedure

Each two-hour session consisted of the following activities. First, all students, at their desks, received a 10-minute introduction—what they would be doing, how it would not effect their math grade, and how it was important that they try their hardest. They were also reminded that they were getting a reward for their participation. Next, all students took a 20-minute pretest. As noted earlier, we created two test forms for the pre- and posttests that were counter-balanced among all students. The tests contained 25 matched items spanning the identified content (geometric sequences) and were administered in paper and pencil format. Calculators were permitted. After the pretest, students were randomly assigned to one of four conditions 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 paper and pencil 10-minute survey.

3. Results

3.1. Feedback and Adaptivity Effects on Learning

We examined Research Questions 1 and 2, addressing 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, from honors to special education) 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 1 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 largest pretest-to-posttest improvement. Confidence intervals (95%) for the posttest data, per condition, are also depicted in the figure.

The general finding of a main effect of condition on learning prompted a specific planned comparison (Bonferroni) involving the three treatment conditions in relation to posttest data. Findings showed that the only significant difference was between the E/A and S/A conditions, where the Mean Difference = 5.62; SE = 2.11; and $p < 0.05$. This suggests that the elaborated feedback feature was primarily responsible for the impact on learning.

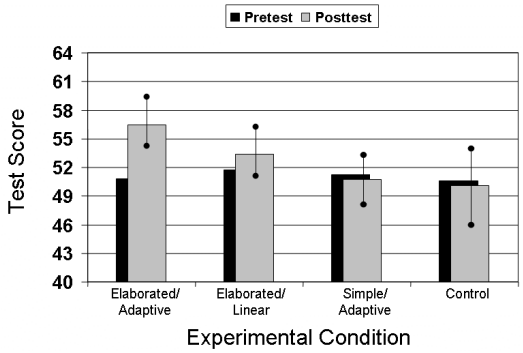

Figure 1. Condition by Pre- and Posttest Scores

3.2. Predictive Validity

The first assessment issue concerns whether the ACED proficiency estimates (relating to the 8 main nodes in the Bayes net) predict outcome performance beyond that predicted by pretest scores. Each proficiency variable possesses a triplet of probabilities, reflecting the estimated probability of being High, Medium, and Low on that proficiency. To reduce the three numbers to a single number, we calculated the Expected A Posteriori (EAP) score: *P*(High) - *P*(Low). The main outcome variable of interest is the EAP for the highest level proficiency in the model—Solve Geometric Sequences (SGS). Higher values of EAP(SGS) should be associated with greater knowledge and skills overall on geometric sequence topics. A regression analysis was computed with posttest score as the dependent variable and (a) pretest score and (b) EAP(SGS) as the independent variables. 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 of the independent variables significantly predicted student outcome data: Multiple $R = 0.71$; $F_{2, 210} = 106.57$; $p < 0.001$. Pretest score and the general estimate of proficiency accounted for 50% of the outcome variance, with EAP(SGS) accounting for 17% of the unique outcome variance over pretest score.

3.3. Reliability of ACED Tasks, Proficiency Estimates, and Outcome Tests

As mentioned earlier, all 63 tasks in the ACED pool of geometric items were statistically linked to relevant proficiencies. Students in all three ACED conditions were required to spend one hour on the computer solving the 63 items. This design decision was made 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 test (via SPSS). The Spearman Brown split-half reliability with unequal halves (i.e., 31 and 32) was equal to 0.84, while Cronbach's $a = 0.88$ which is quite good. Next, we analyzed proficiency estimates from the Bayes net proficiency model, 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. The reliability of the EAP(SGS) score was 0.88. Moreover, the sub-proficiency estimates showed equally impressive reliabilities for their associated tasks suggesting they may be useful for diagnostic purposes. Separate analyses were computed on the reliabilities of the pretest and posttest items. We computed Cronbach's *a* 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. The adjusted pretest *a* = 0.82, and adjusted posttest *a* = 0.85.

3.4. Efficiency of the ACED System

In this study, we required that students in all three ACED conditions spend one hour on the computer solving all 63 items. A typical rationale for using adaptive tests relates to their efficiency capability. That is, adaptive algorithms rely on fewer items to determine proficiency level. The issue here is what the data (proficiency estimates) would look like if we required fewer items to be solved. 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 scores for the highest level node 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 2 shows the results of the plot, confirming our hypothesis.

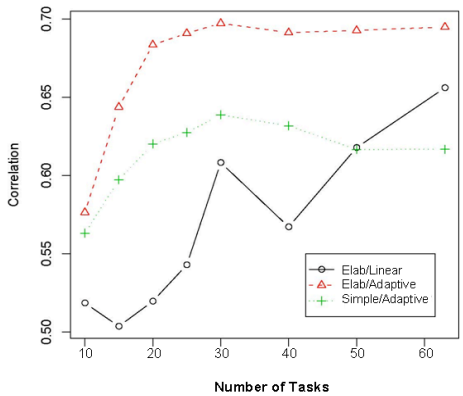

Figure 2. Correlations of EAP (SGS) with Posttest Score by ACED Condition

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 sensible next step would be instructional intervention. In the linear condition (E/L), there is a spike around task 30, then a subsequent drop. When we reviewed the list of the 63 tasks in the order in which they appeared in the linear condition, Tasks 31-36 comprised a set of very difficult items. Finally, the slight decline after Task 30 in the S/A condition could be a fatigue effect. That is, because many of the problems were unfamiliar, and the feedback 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, we note that the noise associated with the correlation estimates is rather large, requiring additional verification of this decline phenomenon.

4. Conclusions

Regarding the learning question, elaborated feedback was, as hypothesized, more effective for student learning than simple feedback (verification only). Nevertheless, the study did not show adaptive sequencing of tasks to be more effective for learning than linear sequencing. However, the adaptive condition did show greater efficiency than the linear condition in achieving high reliability and validity. That is, we saw that for students in the adaptive condition, the ACED assessment could have reasonably terminated after approximately 20 items with no degradation in prediction of outcome. Because most of the students completed all 63 tasks in one hour covering a range of geometric sequence proficiencies (or one task per minute, on average), administering a 20-30 minute test yielding valid and reliable results would be much more efficient—for students and teachers.

Regarding reliability and validity, the ACED system was very reliable in relation to the ACED tasks, proficiency estimates, and pretests and posttests. In terms of validity, based on our findings, it seems that using evidence-based assessment design with Bayes net technology facilitates valid estimation of proficiencies from performance data. Regression analysis showed just a single proficiency estimate—EAP(SGS)—can significantly predict posttest performance, beyond that provided by pretest performance.

In conclusion, we envision a role for the ACED algorithm as part of a larger instructional support system. The adaptive AfL system would continue assessing the student until the best instructional options for that student become clear. Meanwhile, elaborated feedback would ensure that the assessment itself was a valid learning experience.

References

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

[2] Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7-74.

[3] Shute, V. J. (2007). Focus on formative feedback. ETS Research Report, Princeton, NJ.

[4] Stiggins, R. J. (2002). Assessment Crisis: The Absence of Assessment FOR Learning, Phi Delta Kappan Professional Journal, 83(10), 758-765.

[5] Jameson, A. (2006). Adaptive interfaces and agents. In J. A. Jacko & A. Sears (Eds.), Human-computer interaction handbook (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

[6] VanLehn, K. (2006). The behavior of tutoring systems. International Journal of Artificial Intelligence in Education, 16(3), 227-265.

[7] 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.

[8] 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.

[9] Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. Instructional Science, 32, 99-113.

[10] 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.

[11] Narciss, S. & Huth, K. (2004). How to design informative tutoring feedback for multi-media learning. In H. M. Niegemann, D. Leutner & R. Brunken (Ed.), Instructional design for multimedia learning (pp. 181-195). Munster, New York: Waxmann.

[12] 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. New York, ACM Press: 245-252.

[13] 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.

[14] Mason, B. J., & Bruning, R. (2001). Providing feedback in computer-based instruction: What the research tells us. Center for Instructional Innovation, University of Nebraska-Lincoln: 14. Retrieved November 1, 2006, from http://dwb.unl.edu/Edit/MB/MasonBruning.html

[15] Shute, V. J. & Zapata-Rivera, D. (in press). Adaptive technologies. In J. M. Spector, D. Merrill, J. van Merriënboer, & M. Driscoll (Eds.), Handbook of Research on Educational Communications and Technology (3rd Edition). Mahwah, NJ: Erlbaum Associates.

[16] Shute, V. J., Hansen, E. G., & Almond, R. G. (in press). An assessment for learning system called ACED: Designing for learning effectiveness and accessibility, ETS Research Report, Princeton, NJ.

[17] Good, I. J., & Card, W. (1971). The diagnostic process with special reference to errors. Method of Inferential Medicine, 10, 176–188.

[18] 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.

[19] 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.

[20] Conejo, R., Guzman, E., Millán, E., Trella, M., Perez-De-La-Cruz, J. L., & Rios, A. (2004). SIETTE: A web-based tool for adaptive testing. International Journal of Artificial Intelligence in Education 14(1), 29-61.

[21] Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd Edition), Mahwah, NJ: Erlbaum Associates.

[22] 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, and M. Bodvarsson (Eds.). Technology-Based Education: Bringing Researchers and Practitioners Together (pp. 169-202). Greenwich, CT: Information Age Publishing.

[23] Almond, R. G., Yan, D., Matukhin, A, & Chang, D. (2006). StatShop Testing. ETS, Research Memorandum 06-04.