Evaluating a Mixed-Initiative Authoring Environment: Is REDEEM for Real?

Shaaron AINSWORTH and Piers FLEMING

School of Psychology and Learning Sciences Research Institute,
University of Nottingham
Email: {sea/pff}@psychology.nottingham.ac.uk

Abstract. The REDEEM authoring tool allows teachers to create adapted learning environments for their students from existing material. Previous evaluations have shown that under experimental conditions REDEEM can significantly improve learning. The goals of this study were twofold: to explore if REDEEM could improve students' learning in real world situations and to examine if learners can share in the authoring decisions. REDEEM was used to create 10 courses from existing lectures that taught undergraduate statistics. An experimenter performed the content authoring and then created student categories and tutorial strategies that learners chose for themselves. All first-year psychology students were offered the opportunity to learn with REDEEM: 90 used REDEEM at least once but 77 did not. Students also completed a pre-test, 3 attitude questionnaires and their final exam was used as a post-test. Learning with REDEEM was associated with significantly better exam scores, and this remains true even when attempting to control for increased effort or ability of REDEEM users. Students explored a variety of categories and strategies, rating their option to choose this as moderately important. Consequently, whilst there is no direct evidence that allowing students this control enhanced performance, it seems likely that it increased uptake of the system.

1. Introduction

The REDEEM authoring tool was designed to allow teachers significant control over the learning environments with which their students learn. To achieve this goal, the authoring process and the resulting learning environments have both been simplified when compared to more conventional authoring tools. REDEEM uses canned content but delivers it in ways that teachers feel are appropriate to their learners. Specifically, material can be selected for different learners, presented in alternative sequences, with differences exercises and problems, and authors can create tutorial strategies that vary such factors as help, frequency and position of tests and degree of student control. This approach, focussing on *adapted* learning environments rather than *adaptive* learning environments, has been evaluated with respect to both the authors' and learners' experiences (see [1] for a review). Overall, REDEEM was found to be usable by authors with little technological experience and time-efficient for the conversion of existing computer-based training (CBT) into REDEEM learning environments (around 5 hours per hour of instruction). Five experimental studies have contrasted learning with REDEEM to learning with the original CBT in a variety of domains (e.g. Genetics, Computing, Radio Communication) and with a wide range of learners (schoolchildren, adults, students). REDEEM led to an average 30% improvement from pre-test to post-test, whereas CBT increased scores by 23%. This advantage for REDEEM translates into an average effect size of .51, which compares well to non-expert human individual tutors and is around .5 below full-blown ITSs (e.g. [2,3]).

To perform three of these experiments, teachers were recruited who had in-depth knowledge of the topic and the students in this class. They used this knowledge to assign different student categories which resulted in different content and tutorial strategies. In the

other two experiments, this was not possible and all the participants were assigned to one category and strategy. But, it may have been more appropriate to let students choose their own approach to studying the material. This question can be set in the wider context of authoring tools research, namely for any given aspect of the learning environment, who should be making these decisions – should it be a teacher, should it be the system or can some of the authoring decisions be presented to learners in such a way that they can make these decisions for themselves. Whilst, there has been some debate in the literature about how much control to give the author versus the system [4], the issue of how much of the authoring could be performed by learners themselves has received little direct attention. Of course, the general issue of how much control to give students over aspects of their learning has been part of a long and often contentious debate (e.g. [5, 6]). There are claims for enhanced motivation [7] but mixed evidence for the effectiveness of learner control.

However, in the context under consideration (1st year University students), there was no teacher available who could make these decisions based upon personal knowledge of the student. Consequently, to take advantage of REDEEM's ability to offer adapted learning environments, the only sensible route was to allow learners to make these decisions for themselves. As a result, a mixed initiative version of REDEEM was designed that kept the same model of content and interactivity authoring as before, but now gave students the choice of learner category (from predefined categories) and teaching strategy (also predefined). Thus the aim of this approach is not to turn learners into authors as [8] but instead to renegotiate the roles of learners and authors.

A second goal for this research was to explore the effectiveness of REDEEM over extended periods, outside the context of an experiment. One positive aspect of AIED in recent years has been the increase in number of evaluations conducted in realistic contexts (e.g. [3, 9]). However, given the complex issues involved in running an experiment, the norm for evaluation (including the previous REDEEM studies) is that they are conducted in experimental situations with limited curriculum over a short duration and post-tests tend to be on the specific content of the tutor. To show that interacting with a learning environment improves performance when used as part of everyday experience is still far from common (another exception is ANDES [10] whose research goal is to explore if minimally invasive tutoring can improve learning in real world situations). Yet, it is this test that may convince sceptics about the value of ITSs and interactive learning environments. However, assessing if REDEEM improves learning 'for real' is far from easy as it was difficult to predict how many students would chose to use REDEEM or whether we would be able to account for explanations based upon differential use of REDEEM by different types of learners.

2. Brief System Description

REDEEM consists of three components: a courseware catalogue of material created externally to REDEEM, an ITS Shell and a set of authoring tools (please see [1] for a fuller description of components and the authoring process). REDEEM's authoring tools decompose the teaching process into a number of separate components. Essentially, authors are asked to add interactivity to the underlying courseware (by adding questions, hints, answer feedback and reflections points) they describe the structure of material, create student categories and create teaching strategies. This information is then combined by assigning particular teaching strategies and types of material to different learner groups. The difference with this latest version is that the students themselves select one of the learner categories and this now results in a default teaching strategy, which they can change to any other strategies that are available. This design is a trade-off between giving students' significant choice yet only requiring a minimum of interaction to utilise this functionality.

The courseware consisted of ten PowerPoint lectures saved as html. These were then imported into REDEEM by an experimenter, who in addition to describing the structure of the material, added approximately one question per page with an average of three hints per question and an explanation of the correct answer and reflection points. Four learner categories were created (non-confident learner (NCL, confident learner (CL), non-confident reviser (NCR), confident reviser (CR). Four default teaching strategies were created (Table 1) based upon ones teachers had authored in previous studies [11]. In addition, four optional strategies were devised that provided contrasting experiences such as using it in 'exam style' or in 'pre-test' mode (test me after the course, before section or course).

Table 1. Teaching Strategies

Name	**Default**	**Description**
Simple Introduction	NCL	No student control of material or questions; easy/medium questions (max one per page), 2 attempts per question, help available. Questions after page.
Guided Practice	NCR	No student control of material/questions; easy/medium questions (max one per page). 5 attempts per question, help is available. Questions after section.
Guided Discovery	CL	Choice order of sections but not questions. 5 attempts per question, help only on error. Questions after section.
Free Discovery	CR	Choice order of sections and questions. 5 attempts per question, help available
Just Browsing		Complete student control of material. No questions.
Test me after the course		No student control of material or questions. All questions at the end, 1 attempt per question, no help.
Test me before each section		Choose order of sections. Questions are given before each section. 5 attempts per question and help available on error.
Test me before the course		Student control sections All questions at the start. 5 attempts per question. Help is available.

3. Method

3.1. Design and Participants

This study employed a quasi-experimental design as students decided for themselves whether to learn with the REDEEMed lectures. All 215 first-year Psychology students (33 males and 182 females) had previously studied a prerequisite statistics course, which was assessed in the same exam as this course, but for which no REDEEM support had been available. 167 students completed both the pre-test and post-test.

3.2. Materials

Pre and post-tests were multiple-choice, in which each question had one correct and three incorrect answers. A pre-test was created which consisted of 12 multi-choice questions addressing material taught only in the first semester. Questions were selected from an existing pool of exam questions but were not completely representative as they required no calculation (the pre-test was carried out without guaranteed access to calculators). The 100 question multi-choice two hour exam was used as a post-test. These questions were a mix of factual and calculation questions. All students are required to pass this exam before continuing their studies. The experimenters were blind to this exam.

A number of questionnaires were given over the course of the semester to assess students' attitudes to studying, computers, statistics and the perceived value of REDEEM.

- A general questionnaire asked students to report on their computer use and confidence, the amount of time spent studying statistics and the desire for further support.
- An attitude to statistics questionnaire assessed statistics confidence, motivation, knowledge, skill and perceived difficulty on a five-point Likert scale.
- A REDEEM usage questionnaire asked students to report on how much they used REDEEM, to compare it to other study techniques and to rank the importance of various system features (e.g. questions, having a choice of teaching strategy).

3.3. Procedure

- All first year students received traditional statistics teaching for Semester One (ten lectures) from September to December 2003.
- Early in the second semester, during their laboratory classes, students were introduced to REDEEM and instructed in its use. They were informed that data files logging their interactions with the system would be generated and related to their exam performance but data would not passed to statistics lecturers in a way that could identify individuals. During these lessons, students were also given the pre-test and a questionnaire about their use of computers and perceptions of statistics.
- As the second semester progressed, REDEEMed lectures were made available on the School of Psychology intranet after the relevant lecture was given.
- Students logged into REDEEM, chose a lecture and a learner category. Students were free to override the default strategy and change to one of seven others at any time.
- At the end of the lecture course (the tenth lecture) another questionnaire was given to reassess the students' perceptions of statistics and REDEEM.
- Finally, two and a half weeks after the last lecture, all of the students had to complete a statistics exam as part of their course requirements.

4. Results

This study generated a vast amount of data and this paper focuses on a fundamental question, namely whether using REDEEM could be shown to impact upon learning. In order to answer this question a number of preliminary analyses needed to be carried out and criteria set, the most important being what counted as using REDEEM to study a lecture. After examining the raw data, it was concluded that a fair criterion was to say that students were considered to have studied a lecture with REDEEM if they had completed 70% of the questions for that lecture. The range of strategies allowed very different patterns of interactions, so questions answered was chosen because many students only accessed the practice questions without choosing to review the material and only one student looked at more than three pages without answering a question. Note, this criterion excludes the just browsing strategy, but this was almost never used and was no one's preferred strategy.

A second important preliminary analysis was to relate the 100 item exam to individual lectures. This was relatively simple given the relationship between the exam structure and learning objectives set by the lecturers. 42 questions were judged as assessing Semester 1 performance and so these questions provided a score on the exam that was unaffected by REDEEM. The Semester 2 questions were categorised according to the lecture in which the correct answer was covered. The 12 questions that addressed material taught in both semesters were not analysed further.

4.1. Relationship between REDEEM Use and Learning Outcomes

Table 2. Scores of REDEEM v non-REDEEM users

	Pre-test	Semester 1 Post-test	Semester 2 Post-test
REDEEM at least once (N = 90)	50.64% (15.96)	69.00% (12.08)	58.09% (13.03)
Never used REDEEM (N = 77)	49.24% (14.06)	67.32% (10.35)	53.44% (14.43)

The first analysis compared the scores of students who had never used REDEEM to those who had studied at least one lesson with REDEEM (Table 2). A [2 by 1] MANOVA on the pre-test, Semester 1 and Semester 2 scores revealed no difference for pre-test and Semester 1, but found the REDEEM users scored higher on Semester 2 ($F(1,167) = 4.78$, $p<.03$). However, this simple contrast overlooks much of the subtlety of the data. Of the 10 lectures; some students studied only 1 or 2 lectures and some all 10. Hence, the amount of REDEEM use (no. of lectures completed to 70% criterion) was correlated with exam scores (Table 3) - the more lectures studied with REDEEM, the greater the Semester 2 scores.

Table 3. Correlation between Test Scores and REDEEM use

	Pre-test scores	Semester 1 score	Semester 2 score	No. of lectures
Pre-test score		.171*	.165*	.038
Semester 1 score			.436***	.116
Semester 2 score				.287***
No. of lectures				

Note. * = p<.05, ** = p<.01, *** = p<.001 (two tailed test of significance)

A stepwise linear regression predicted the influence of REDEEM use and Semester 1 performance on Semester 2 performance. Semester 1 performance and REDEEM use combined predicted 23.7% of the variance (adjusted R squared). The model was significant ($F(2, 164) = 26.814$, $p<.001$). Beta values show that semester 1 performance (Beta = 0.415, $t = 6.097$, $p<.001$) is approximately twice as important as REDEEM use (Beta = 0.238, $t = 3.50$, $p< .001$) but both were significant predictors. Participants were predicted to do about 1% (exactly 0.954%) better for each REDEEM lecture they completed.

These analyses suggest that REDEEM improves students' performance, but it is still possible to argue that those students who used REDEEM more frequently were harder working and motivated students. A stringent test of the effectiveness of learning with REDEEM was to examine each lecture's questions on the exam individually. Furthermore, Semester 1 scores provide a good control for enhanced effort or ability. Consequently, ten ANCOVAS (partialling out Semester 1 performance) compared performance between REDEEM users (for that lecture) and non-REDEEM users (for that lecture). Performance for lectures 4, 5, 7 and 8 was significantly better for REDEEM users ($F(1,179) = 9.34$, $p<.003$; $F(1,179) = 4.36$, $p<.04$; $F(1,179) = 4.26$, $p<.04$; $F(1,179) = 8.94$, $p<.01$) (**Table 4**).

Table 4. Percentage Scores and the Number of the Questions on the Exam by Lecture

Lect.	No Ques	REDEEM users	REDEEM Non-users	Lect.	No Ques	REDEEM users	REDEEM Non-users
1	3	79.17 (23.47) N = 78	72.49 (27.33) N = 104	6	2	80.85 (30.49) N = 47	68.89 (34.48) N = 135
2	2	68.75 (33.92) N = 64	69.07 (35.16) N = 118	7	9	51.83 (19.83) N = 48	41.99 (22.54) N = 134
3	7	58.42 (19.83) N = 55	56.30 (19.47) N = 127	8	4	75.56 (25.83) N = 45	58.03 (30.61) N = 137
4	6	73.51 (22.84) N = 63	61.42 (24.90) N = 119	9	3	30.93 (30.39) N = 43	29.88 (24.25) N = 139
5	9	56.08 (16.62) N = 48	49.19 (18.49) N = 134	10	1	60.53 (49.54) N = 38	59.03 (49.35) N = 144

4.2. Student's use of REDEEM and their Perceptions of the Features Helpfulness

Participants completed questionnaires about their attitude to and experiences of computing and statistics. Consequently, we can explore if this influenced REDEEM uptake. No measure of statistical confidence, motivation or perception of statistics difficulty was related to REDEEM use (all correlations were between .11 and -.10). Similarly, no amount of computer usage or confidence influenced REDEEM usage.

Table 5. Most Commonly Chosen Category and Strategy (with Default Strategy)

Category	% Choice	Strategy	% Choice
Non confident learner	20.9%	Simple introduction (default)	21.3%
Non confident reviser	32.6%	Guided practice (default)	28.7%
Confident learner	7.0%	Guided discovery (default)	6.4%
Confident reviser	39.5%	Free discovery (default)	8.5%
		Test me before each section	6.4%
		Test me after the course	9.6%
		Test me before the course	19.1%

Students choose a learner category (and teaching strategy) for each lecture (Table 5). The choice of categories is not equal; very few students chose the category of "Confident learner". Partly as a result, few students experienced the Guided Discovery strategy. In terms of strategy, it is notable that "Confident revisers" were most likely to explore other strategies, and in particular to select "Test me before the course".

Table 6. Students who Chose Confident versus Non-Confident Categories

	Non-Confident (N=56)	Confident (N=32)
Pre-test	47.62% (15.94)	54.68% (14.65)
Semester 1	68.33% (12.21)	70.28% (9.86)
Semester 2	56.68% (12.17)	60.73% (14.37)
Confidence	1.80 (0.79)	2.37 (0.79)
Knowledge	1.98 (0.83)	2.50 (0.88)
Difficulty	1.64 (0.75)	2.28 (0.95)
Motivation	2.64 (0.99)	2.63 (1.01)

NB 2 subjects did not complete all parts of the statistics attitude questionnaire

We further analyzed whether the confidence levels expressed on statistics attitudes questionnaire related to student choice of learner category and to measures of learning. Whilst there was no relationship between reviser/learner and performance measures or attitude, but there was for confident/non-confident categories. Analysis by MANOVA (see Table 6) found that students who chose non-confident categories also judged themselves as less confident and knowledgeable on the statistics attitude questionnaire and found statistics more difficult than other subjects ($F(1,86) = 9.17$, $p<.003$; $F(1,86) = 9.22$, $p<.003$, $F(1,86) = 12.03$, $p<.001$). They also differed on their pre-test scores ($F(1,86) = 4.23$, $p<.043$) but not at post-test nor in their motivation to learn statistics.

Those students who attended the final lecture and completed the questionnaire (99 students, 60 of whom reported using REDEEM) considered REDEEM to be less useful than lectures for learning statistics, but more useful than tutorials, textbooks or working with friends. They reported they would definitely use REDEEM for the 2nd year statistics course if it was available (4.47/5) and would recommend REDEEM to next year's first years (4.45/5). Only one respondent would not use or recommend REDEEM. They ranked REDEEM's features in the following order of usefulness: Questions, Hints & Explanations of Answers, Choice of Strategy, Review facilities, Student History, and Notes tool.

5. Discussion

A number of analyses were performed to examine if use of REDEEM could be shown to impact upon exam performance. No analysis found that use of REDEEM was associated with either higher pre-test scores or higher Semester 1 scores on the post-test. However, Semester 2 scores were influenced by use of REDEEM. Those students who used REDEEM performed better and the more they studied with REDEEM the better they did. Furthermore, students whose used REDEEM more still performed better even with Semester 1 scores partialled out, which mitigates against an explanation based solely on differential use of REDEEM by motivated or higher ability students. Finally, REDEEM increased performance specifically on the exam questions that corresponded to the lectures that a student had studied with REDEEM (an average of 64% on those lecture's questions versus 54% on those they had not studied). These analyses combine to suggest that studying with REDEEM enhanced performance (1% per lecture studied), which if students complete all 10 lectures would result in an improvement of a degree class in the UK system.

Students who chose to learn with REDEEM did not differ in their attitudes to statistics, prior knowledge of statistics or attitudes to and use of computers. Studying with REDEEM does not seem to differentially attract students with different characteristics. However, their views about statistics did influence the way they used REDEEM. Students who rated themselves as less statistically confident tended to also choose the non-confident student category and also tended to have lower pre-test scores. Consequently, it would appear that students' lack of confidence about their statistics knowledge was rooted, to some extent, in insight into their understanding of statistics at the beginning of Semester 2. However, by the end of the year, these students had the same exam scores as their peers.

Students rated learning with REDEEM fairly highly. They did not see it as a substitute for lectures (nor was it intended to be), but preferred REDEEM to other forms of independent study. No doubt this was related to the provision of questions (with help and explanations) given the high rating of this feature and students' view that the statistics course should be supplemented with more questions (81%).

The second goal of this study was to explore if sharing some of the authoring decisions between student and author was helpful, consequently this version of REDEEM allowed students to choose their own learner category and strategy from pre-defined author choices. Students tended to pick revision categories rather than learning categories. This is almost certainly related to the fact that approximately 2/3rd of REDEEM use occurred after the end of term and a stunning 25% of total use was in the 36 hours prior to the exam. This also helps to explain the gradual fall in REDEEM use across lectures – many students simply started at lecture 1 and ran out of time to complete the whole course. Whilst this may not be an ideal way to learn statistics, it does show that REDEEM can provide support for students at times in which the traditional university provision is unavailable. Students were more equally split between those who chose to learn as either confident or non-confident. This choice was consistent both with their attitude to statistics and with poorer performance at the pre-test for non-confident learners.

For this study, there was no difference for the alternative categories in the sequence and structure of material (it simply replicated the original lecture), but each category had a different default teaching strategy based on previous experimental studies. Most of the students stuck with the default strategy except "Confident revisers" who rarely used the default strategy. This may indicate that students in this category had the confidence to explore the range of tutorial strategies or may also indicate that the default strategy was not appropriate. Many swapped to "Test me before the course", which is particularly interesting as generally this strategy was rated as the second least useful (4.53/7 compared to the most valued "Test me after the course" 6.55/7). This apparent contradiction can be resolved by

examination of the dates of the log files which showed it was this strategy that was used increasingly in the days before the exam. This suggests that a category of "Last minute reviser" with a "Test me before the course" strategy may be a useful future option.

This study cannot reveal the contribution that choosing categories or strategies played in improving learning outcomes or enhancing uptake of this system. Nor can we determine the appropriateness of our decisions about learner categories or strategies. Students' choice of categories does seem highly rational given the relationship to statistics attitudes and prior knowledge, and the time of year when use occurred. Students rated their opportunity to choose teaching strategies as the next most important feature after REDEEM's question features (4.45/7). If we had used the previous version of REDEEM where authors chose the strategy it is likely we would have picked a strategy most like "Guided Discovery". Overall, this was the least used strategy after "Just Browsing" because students rarely chose the "Confident Learner" strategy. Again we have no way of ascertaining if our choice or student's individual choices would have resulted in better learning outcomes, but it is probable that this strategy would not have suited their last minute revision tactic.

Analysis of the data is on-going to explore how to improve the authoring of such features as questions and hints (e.g. why did studying lecture 3 not improve performance?) as well improvements to choices offered for learner category and teaching strategy. However, experiments with controlled use of the system and this quasi-experimental study suggest that learning with REDEEM is more helpful than learning without it.

6. References

[1] S. E. Ainsworth, N. Major, S. K. Grimshaw, M. Hayes, J. D. Underwood, B. Williams, and D. J. Wood, "REDEEM: Simple Intelligent Tutoring Systems From Usable Tools," in *Tools for Advanced Technology Learning Environments.*, T. Murray, S. Blessing, and S. E. Ainsworth, Eds. Amsterdam: Kluwer Academic Publishers, 2003, pp. 205-232.

[2] A. C. Graesser, N. K. Person, D. Harter, and T. T. R. Group, "Teaching Tactics and Dialog in AutoTutor," *International Journal of Artificial Intelligence in Education*, vol. 12, pp. 257-279, 2001.

[3] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark, "Intelligent tutoring goes to school in the big city," *International Journal of Artificial Intelligence in Education*, vol. 8, pp. 30-43, 1997.

[4] T. Murray, "Authoring intelligent tutoring systems: An analysis of the state of the art.," *International Journal of Artificial Intelligence in Education*, vol. 10, pp. 98-129, 1999.

[5] V. Aleven and K. R. Koedinger, "Limitations of student control: Do students know when they need help?," in *Intelligent Tutoring Systems: Proceedings of the 5th International Conference ITS 2000*, vol. 1839, *Lecture Notes in Computer Science*, 2000, pp. 292-303.

[6] T. C. Reeves, "Pseudoscience in computer-based instruction: The case of learner control literature," *Journal of Computer-Based Instruction*, vol. 20, pp. 39-46, 1993.

[7] T. W. Malone, "Toward a theory of intrinsically motivating instruction," *Cognitive Science*, vol. 5, pp. 333-369, 1981.

[8] I. Arroyo and B. P. Woolf, "Students in AWE: changing their role from consumers to producers of ITS content," in *Advanced Technologies for Mathematics Education Workshop. Supplementary Proceedings of the 11th International Conference on Artificial Intelligence in Education.*, 2003.

[9] P. Suraweera and A. Mitrovic, "An Intelligent Tutoring System for Entity Relationship Modeling," *International Journal of Artificial Intelligence in Education*, vol. 14, pp. 375-417, 2004.

[10] K. VanLehn, C. Lynch, L. Taylor, A. Weinstein, R. Shelgy, K. Schulze, D. Treacy, and M. Wintersgill, "Minimally invasive tutoring of complex physics problem solving," in *Proceedings of the 6th International Conference ITS 2002*, vol. 2363, S. A. Cerri, G. Gouardères, and F. Paraguaçu, Eds. Berlin: Springer-Verlag, 2002, pp. 367-376.

[11] S. E. Ainsworth and S. K. Grimshaw, "Evaluating the REDEEM authoring tool: Can teachers create effective learning environments," *International Journal of Artificial Intelligence in Education*, vol. 14, pp. 279-312, 2004.