

# Comparing the Learning Effectiveness of REDEEM and CBT

Shaaron AINSWORTH, Ben WILLIAMS and David WOOD

*School of Psychology, University of Nottingham, Nottingham, NG7 2RD, UK*

*Email sea@psychology.nottingham.ac.uk*

**Abstract.** The REDEEM authoring tool creates simple ITSs from existing computer-based training (CBT). In this study, learning outcomes of 33 adult volunteers, who studied 'Introduction to Personal Computing and Networking' courses with CBT and with REDEEM, were compared. Performance of all students improved from pre-test to post-test and was significantly better on the course students had learnt with REDEEM. Analysis of students' performance suggested that this REDEEM enhancement was due to the increased interaction that REDEEM promotes. Those students who learnt most from REDEEM tended to be those who experienced REDEEM first and who required less attempts to answer REDEEM's questions correctly. This study showed how simple improvements to educational software can significantly increase students' learning.

**Keywords:** Evaluation, Macroadaptation, ITS Authoring Tools

## 1. Introduction

As the field of ITS Authoring Tools (ITSATs) matures, more emphasis is being placed on evaluating the success of these systems. Initial evaluations focussed on the usability of these ITSATS and considered whether authors could create their desired ITSs with these tools in a time and cost effective manner [1, 2, 3]. This focus on ease of use has not been uncontentious as there is considerable debate (*e.g.* [4]) about whether for an ITSAT to be usable by non-programmers, it is worth sacrificing some of the resulting ITS's flexibility and power. Two ITSATS, which as they have evolved have increasingly emphasised ease of use, are REDEEM and XAIDA. REDEEM [5] is designed for use by classroom teachers and creates ITSs in any domain but without a runnable domain model. XAIDA [6] builds tutors in areas such as maintenance training, algebra, biology and computer literacy. It is designed for use by non-experts and has relatively shallow knowledge of what it is teaching. It is not known if the compromise between power and usability that these systems represent is productive, especially as previous research on ITSs has tended to show that "more intelligence" increases learning outcomes (*e.g.* [7, 8, 9]). Thus, the question remains, can ITSATS be used by non-experts to create effective learning experiences for students?

The "gold standard" comparison for learning effectiveness is often considered to be that of an expert human tutor. Bloom [10] argues that one-to-one tutoring by expert tutors produced an average gain in test scores of two standard deviations (a 2 sigma effect) compared to traditional whole class teaching. Non-experts are not quite as effective but can still improve tutoring by around of 0.4 sigmas [11]. Evaluations of ITSs reveal effect sizes of between 0.4 and 1 sigmas compared to classroom teaching (*e.g.* AUTOTUTOR, [12]; CMU algebra tutors, [13]). Hsieh et al [6] review 13 unpublished studies with XAIDA which show that students understanding of the domain improved. A small number of these

studies compared XAIDA to other forms of instruction. Wenzel *et al* (cited in [6]) found equal improvements in learning for students studying from XAIDA and from a lecture. Cascaus *et al* (in [6]) compared practice with adaptive and non-adaptive versions of XAIDA and found overall improvements in learning of which a little could be explained by increased adaptation. These studies do not report an effect size but given the general lack of significant difference between conditions, it is unlikely to be large. They show that XAIDA can be used to create effective ITSs but say little about their relative efficacy compared to other forms of instruction. The study reported in this paper is an experimental evaluation of the effectiveness of the REDEEM Authoring Environment when compared to more traditional CBT. Before describing the results of this evaluation, we will briefly describe the REDEEM system and previous evaluation studies.

## 2. System Description

REDEEM allows authors to import existing computer-based training as domain content and then use the Tools to overlay their teaching expertise. The REDEEM Shell uses this knowledge, together with its own default teaching knowledge, to deliver the courseware adaptively to meet the needs of different learners. REDEEM therefore consists of three components: a courseware catalogue of material created externally to REDEEM, a set of authoring tools and an ITS Shell. There are essentially two stages to REDEEM authoring; in the first stage the domain material is enriched but it remains essentially non-adaptive CBT; in the second stage the CBT is individualized to the needs of learners by macro-adapting the teaching strategies and the content to student categories.

Authors first select courseware, which consists of individual pages of material containing text, animations, simulations, *etc*, in either Toolbook or html format. Then authors use the tools to provide a domain model of this material by describing its characteristics (*e.g.* by grouping into sections, and describing the content of pages and sections), which allows the ITS Shell to sequence and structure it. To increase interactivity, authors can associate a reflection point or non-computer task with a page. They can also create questions (such as multiple-choice, fill-in-the-blank, and true-false) and provide feedback that explains why an answer is correct. In addition, REDEEM aims to offer multiple levels of help in way that is similar to contingent help [14], so authors can create up to five hints for each question, which ideally increase in specificity. Finally authors describe a number of question characteristics (*e.g.* difficulty, pre-test or post-test) that the Shell uses in combination with a teaching strategy to decide how to ask each question.

Then authors individualize a course to meet the requirements of their specific learners. Different teaching strategies are created by manipulating dimensional sliders of eight components of instruction (*e.g.* student control, position and amount of question, help, number of attempts at questions) and by including appropriate questions. Authors classify students into categories (based upon any dimension they choose), which can dynamically adjust if the categories are based on performance measures. These categories are associated with a teaching strategy and the author's choice of content.

## 3. Previous Evaluations

Previous studies with REDEEM have shown that it is one of the most usable ITS authoring tools. It takes only 90 minutes to train authors and the majority of the tools are simple to use. Authors average one to four hours to develop an hour of instruction (see [2,15]). Adding interactivity is relatively time-intensive and varies considerably between authors, however, individualising ITSs to specific learners' needs rarely takes more than 30 minutes.

The question of whether REDEEM could be used to create effective learning experiences was addressed by [16]. A secondary school teacher was given two previously developed courses that teach the age 14-16 UK curriculum on Genetics. These formed the CBT used as the basis for comparison. She was then asked to author her ideal ITSs from this CBT for the 86 14-16 yr old pupils that she taught. Compared to the CBT, the resulting REDEEM ITSs were significantly more interactive with nearly 75% of pages having an associated question and up to five hints per question as well as reflection points and non-computer tasks. She then macro-adapted REDEEM to five ability categories such that each category received a different teaching strategy and content. Using a crossover design, the learning outcomes for students who studied with REDEEM and CBT were then compared. Participants improved their knowledge of genetics but their improvement was the same whether they received a course as CBT (an average improvement of 8.2%) or REDEEM (10.4%). This was true for learners in all student categories and so with all versions of REDEEM. However, learners who engaged more with REDEEM by writing more notes, spending longer interacting with REDEEM and who answered more questions correctly first time did learn significantly more.

This study appeared to show that enriching and macroadapting CBT with REDEEM does not lead to improved learning. However, it is unclear if this result was due to the authoring decisions, the Shell's interpretation of these decisions, the external courseware used or issues concerning the study's implementation. Hence, this paper reports a study with a similar design but run a more tightly controlled way. We recruited with adult volunteers for shorter courses and used researcher-developed ITSs.

#### 4. Authoring Phase

The courseware used in this experiment was based on two courses developed by the Royal Naval School of Education and Training Technology for use in Naval colleges. The first course provided an introduction to personal computing and was 60 pages long and the second course introduced working with networks. The courses consist of text and graphics declarative material with some multimedia, animation and simple exercises. Navigation through the course is linear. They were adapted for use in this study by removing Navy specific material and by extending the courses to include some more difficult concepts.

A researcher then used the REDEEM Tools to create two simple REDEEM ITSs (basing his authoring on that done initially by a Naval trainer). In contrast to [16], the REDEEM differentiation features were not used, as we had no personal knowledge of the students in the classes. This resulted in one learner category; so all students saw the same material and received the same teaching strategy. Thus, the main difference between the two courses is in the interactivity supported by REDEEM rather than in the use of individualisation. As such both the ITSs and the experimental design are considerably simplified compared to [16]. The main authoring decisions and, in consequence, the additional features in the REDEEM courses are summarised in Table 1.

**Table 1. Comparison of REDEEM and CBT**

	REDEEM Features
Content	REDEEM and CBT provide the same content but REDEEM provides section introductions and summarises progress.
Level of Control	Author controlled, so similar to CBT.
Questions	20 questions per course. Questions asked at the end of a section. Students allowed multiple attempts to answer question correctly. None of these questions in CBT
Help	On request and error with between 1 and 5 hints per question.
Reflection points	11 per course in REDEEM. General advice to make notes in CBT.

## 5. Method

### 5.1 Participants

Twenty men and thirteen women participated in the study. All worked at RAF Waddington and had responded to advertisements offering them the chance to improve their knowledge of computers.

### 5.2 Material

A 60 item multiple-choice test was developed. It consisted of 30 questions on the Personal Computing and 30 on Networking. Questions for each course were further subdivided into the 20 questions that were asked during the REDEEM intervention and 10 questions that addressed material covered by the course but were not directly questioned by REDEEM.

### 5.3 Design

A crossover design was used such that all participants received one course under REDEEM and one as CBT only, *i.e.* half received REDEEM Personal Computing (PC) and CBT Networking (Network) and half CBT PC and REDEEM Network.

### 5.4 Procedure

1. A Raven's Progressive Matrices test was given to all participants.
2. Participants took the Personal Computing test and then were randomly assigned to study either the REDEEM or CBT PC Course.
3. Immediately following the intervention, participants completed the PC test again.
4. After a short break, the same procedure was followed with the Networking Course.
5. Upon completion of both courses, students were asked to complete a short questionnaire that asked them to evaluate their experience of learning with the two systems.

## 6. Results

### 6.1. Learning Outcomes

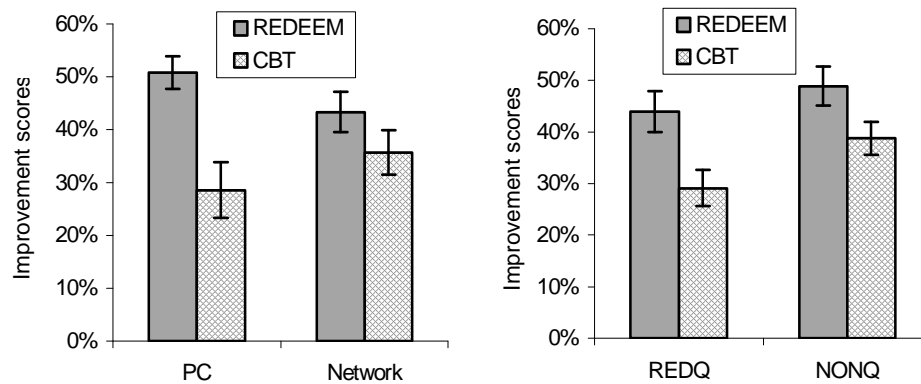
To examine the effects of the intervention, a [2 by 2 by 2] mixed design ANOVA was carried out on the pre-test and post-test data. The design of the analysis was 2(PC, Network) by 2(pre-test, post-test) with a between subjects factor of order of environment (REDEEM PC/CBT Net, REDEEM Net/CBT PC).

**Table 2.** Pre and Post Test Percentage Scores by Course and Environment

	REDEEM				CBT			
	PC (n = 16)		Network (n = 17)		PC (n = 17)		Network (n = 16)	
	Mean	S D	Mean	S D	Mean	S D	Mean	S D
Pre-test	28.1%	11.2	25.8%	8.2	24.0%	9.3	25.0%	7.3
Post-test	78.5%	13.3	70.7%	15.5	52.5%	18.4	63.5%	15.8

Analysis revealed a significant main effect of time ( $F_{1,31} = 237.03$ ,  $MSE = 218.0$ ,  $p < 0.0001$ ), with subjects scoring higher at post-test. This was modified by significant interaction between course and environment ( $F_{1,31} = 64.93$ ,  $MSE = 40.0$ ,  $p < 0.0001$ ) and a three way significant interaction between time, course and environment ( $F_{1,31} = 29.81$ ,  $MSE$

= 61.58,  $p < 0.0001$ ). To simplify the analysis, gain scores per subject were calculated (post-test – pre-test) and analysed using a [2 by 2] mixed ANOVA with environment and course as factors (graphed in Figure 1). This showed a single significant interaction ( $F_{1,31} = 29.91$ ,  $MSE = 123.14$ ,  $p < 0.0001$ ). Simple main effects analysis confirmed that subjects improved their scores more on the course they experienced through REDEEM ( $F_{1,31} = 14.2$ ,  $MSE = 123.1$ ,  $p < 0.005$  and  $F_{1,31} = 15.8$ ,  $MSE = 123.1$ ,  $p < 0.005$  for CBT/RED and RED/CBT respectively). However, the difference between REDEEM and CBT was only significant for the PC course ( $F_{1,62} = 14.6$ ,  $MSE = 279.6$ ,  $p < 0.005$ ).



**Figure 1.** Improvement Scores by Environment and Course

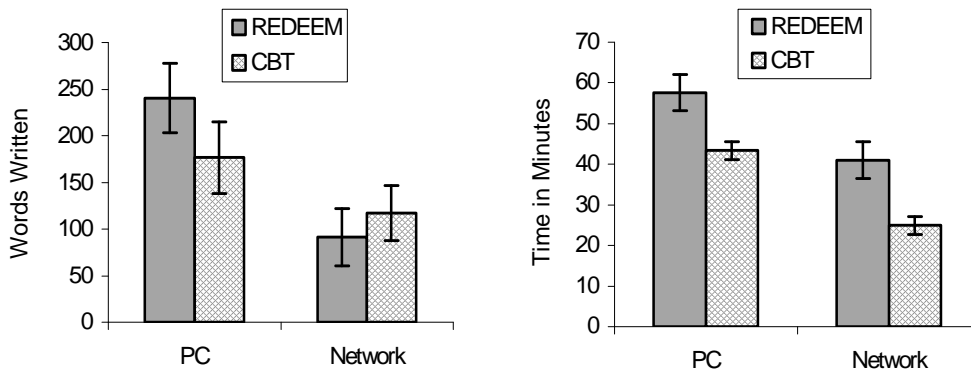
**Figure 2.** Improvement Scores by Environment and Question Type Collapsed Across Course.

It would be expected that the benefits of REDEEM would manifest most strongly on test items that had formed part of the intervention. This was examined using a [2 by 3 by 2] ANOVA on improvement scores. The design of the analysis was 2(PC, Network) by 2(Redeem questions (RedQ), non Redeem questions (NonQ)) with a between subjects factor of order of environments (REDEEM or CBT first). Analysis revealed a significant main effect of environment ( $F_{1,31} = 18.21$ ,  $MSE = 8.70$ ,  $p < 0.001$ ), with participants improving more on the questions that corresponded to their REDEEM course. This was modified by significant interaction between environment and question type ( $F_{1,31} = 6.84$ ,  $MSE = 2.50$ ,  $p < 0.02$ ) (Figure 2). Simple main effects analysis showed there was an effect of environment on RedQ questions ( $F_{1,62} = 24.16$ ,  $MSE = 2.70$ ,  $p < 0.0001$ ), *i.e.* improvement was greater on REDQ questions for the course experienced through REDEEM rather than as CBT. However, contrary to prediction, there was also a difference between RedQ and NonQ questions for the CBT group with NonR questions associated with enhanced performance ( $F_{1,62} = 5.84$ ,  $MSE = 2.80$ ,  $p < 0.05$ ).

### 6.3 REDEEM and CBT process measures

Two types of process measure can be calculated for both REDEEM and CBT – the number of notes written and the amount of time spent learning. To calculate the notes written, both words written on paper and in REDEEM were totalled. Analysis by [2 by 2] mixed ANOVA with factors of course and environment showed a single significant effect – more notes were written on the PC course ( $F_{1,31} = 55.25$ ,  $MSE = 9530.14$ ,  $p < 0.0001$ ) (see Figure 3). Number of notes written does not correlate with any measure of performance (pre-test, post-test, improvement or Raven's score). Analysis of the time data showed that students spent considerably longer on the PC course ( $F_{1,28} = 28.86$ ,  $MSE = 258.9$ ,  $p < 0.001$ ) and there was also a significant interaction between course and environment ( $F_{1,28} = 38.31$ ,  $MSE = 119.8$ ,  $p < 0.001$ ). Simple main effects analysis showed that students spent longer working with REDEEM on both the PC and Networking courses ( $F_{1,56} = 10.9$ ,  $MSE =$

188.9  $p < 0.002$  and  $F_{1,56} = 13.46$ , 188.9  $p < 0.002$  respectively). However, those subjects who learned with REDEEM first spent considerably less time on the following CBT package ( $F_{1,28} = 66.8$ ,  $MSE = 119.7$ ,  $p < 0.001$ ) whereas there was no difference between time spent with the different environments for those who began with CBT and then progressed to REDEEM. Time spent learning with either environment does not correlate with any measure of performance (pre-test, post-test, improvement or Raven's score).



**Figure 3 and 4.** Number of notes written (3) and time spent learning (4) by environment and course

### 6.4 REDEEM Process Data

REDEEM logs information about how many attempts a student requires to answer the question correctly and the number of hints either provided or requested as part of the report given to teachers. These measures were analysed to examine if there was any systematic relationship between behaviour with REDEEM and performance, prior knowledge or Raven's scores. Data is presented for the learner's REDEEM interaction, irrespective of whether it was the PC or Networking course.

**Table 3.** Correlation between Raven's scores, learning gains, hints requested, and attempts at the question.

	2	3	4	5
Raven's score	0.04	0.14	-.30*	-.46***
REDEEM pre-test score		-0.35**	-0.07	-.16
REDEEM Gain score			-0.01	-.38**
Hints on Requests				0.35**
Attempts at Question				

Note \* =  $p < 0.1$ , \*\* =  $p < 0.05$ , \*\*\* =  $p < 0.01$  (two tailed test of significance).

Students who learnt more with REDEEM tended to know less to start with but they required fewer attempts to answer a question correctly. There was no relationship between the number of hints requested and learning outcomes, although requesting more hints was associated with increased attempts at questions. Pre-test does not predict those students who required more attempts to answer questions correctly, however Raven's scores did, with those students with lower scores requiring more attempts.

## 7. Discussion

Students' performance at post-test was significantly higher than it was at pre-test. Furthermore, learners' scores differentially improved for the course they learnt with REDEEM. This difference is educationally as well as statistically significant. An effect size analysis showed that REDEEM led to an average 0.81 sigma improvement in learning

outcomes compared to CBT (if calculated separately REDEEM PC led to a 0.96 sigma improvement and REDEEM Network 0.71 compared to the equivalent CBT). Accordingly, REDEEM ITSs do not deliver the learning gains associated with expert human tutors but they are in the same range as traditional ITSs and non-expert tutors compared to classroom teaching. The difference between REDEEM and CBT was not as great on the Networking course as on the PC course. This is plausibly an order effect as there are no obvious differences in authoring between the two courses. Students transferring from REDEEM to CBT Network to do slightly better than predicted and those transferring from CBT to REDEEM Network slightly worse. This is particularly noticeable as students spent less time learning on CBT Network. They also consistently write more notes on the PC course (irrespective of environment). A planned full cross over design (*i.e.* two further conditions of REDEEM/ REDEEM and CBT/ CBT) will provide more insight into these effects.

One of the key differences between the REDEEM ITSs and the CBT is that REDEEM asks students questions, providing hints to their solution and explaining why answers are correct. Hence, performance should be most facilitated on those items that REDEEM directly questions. Therefore, we hypothesised that learning should be enhanced for REDEEM questions compared to Non REDEEM questions on the REDEEM course and there would be no difference between the question types on the CBT course. However, whilst the data did show the predicted benefits of REDEEM, they did not manifest in this way. Firstly, the results showed that non REDEEM questions were associated with enhanced performance for the CBT course, *i.e.* when no questions are asked, material that formed part of the NONQ battery was more easily learnt than material that formed part of the REDQ battery. Hence, REDEEM improves learning by removing the NONQ over REDQ advantage on the student's REDEEM course as students' performance is better on the REDQs that they learnt with REDEEM rather than the ones they learnt with the CBT. We suggest that the authoring on these courses was influenced by an unconscious bias to set students questions on items that authors expected to be more difficult. In this way, the authors hoped to provide students with more opportunities to learn difficult material and given the REDEEM advantage, it would appear that they have been successful.

The results of this study indicated that some students were likely to learn more with REDEEM than others. They tended to start with lower pre-test scores, although a simple ceiling effect can be ruled out as the average post-test scores was 73%. They were less likely to need multiple attempts in order to answer questions correctly. Furthermore, students who had required fewer attempts at questions had higher Raven's scores (in line with claims that Raven's measures an educative component of general intelligence). Students who needed multiple attempts also asked for more help, although this is explained by learners requesting help after receiving an incorrect error message. The results of these analyses will help provide information concerning what factors we should take into account when using REDEEM macro-adaptive features to adjust teaching strategies to the needs of particular students. They also indicate features that REDEEM could monitor to adjust its behaviour or inform teachers of potentially struggling students. For example, increased number of attempts needed to answer question correctly was also associated with poorer learning outcomes in [16] with younger students studying different material.

After completing the courses, the participants filled in a questionnaire about their experience. All but one of the participants said they preferred learning with REDEEM. Their explanations as to this preference again tended to focus on REDEEM's questions (*e.g.* "questions helped me check I had understood before leaving a section"), with hints and explanations of answers, and section summaries also being identified as useful.

In summary, the REDEEM ITSs used in this study were enriched with extra interactive elements compared to the CBT and we endeavoured to apply an appropriate teaching strategy. However, the extra "intelligence" that REDEEM can provide such as the

use of student stereotypes, multiple teaching strategies, differentiation of content and monitoring of student performance was not used. Even without these features, learning gains with REDEEM were both statistically and educationally greater. This may suggest that some of the previous comparisons of ITSs to lectures and classroom teaching found enhanced learning because of the increased interactivity associated with ITSs rather than their ability to micro or macro adapt. However, further research with REDEEM will help us uncover whether adding more micro-adaptive functions to the REDEEM shell and utilizing REDEEM's macro-adaptive features will provide further improvement upon this base line.

## 8. Acknowledgements

This research was supported by the Office of Naval Research under grant number N000 14-99-1-0777 at the ESRC Centre for Research in Development, Instruction and Training. We would like to thank Alan Richardson and PLA John Allison for their help in running the study, RNSETT for providing the courseware, CPO Pete Skyrme for help with the authoring and all the volunteers at RAF Waddington for participating so enthusiastically.

## 9. References

1. Bell, B., *Supporting Educational Software Design with Knowledge-Rich Tools*. International Journal of Artificial Intelligence in Education, 1999. **10**: p. 46-74.
2. Ainsworth, S.E., J. Underwood, and S. Grimshaw, *Formatively evaluating REDEEM - An authoring environment for ITSs*, in *Proceedings of the 10<sup>th</sup> International Conference on AI in Education*, 1999. p. 93-100.
3. Murray, T., *Authoring intelligent tutoring systems: An analysis of the state of the art*. International Journal of Artificial Intelligence in Education, 1999. **10**: p. 98-129.
4. Murray, T., *Having it All, Maybe: Design Tradeoffs in ITS Authoring Tools.*, in *Proceedings of the Third International Conference on Intelligent Tutoring Systems*, 1996, Springer-Verlag: Berlin. p. 93-101.
5. Major, N., S.E. Ainsworth, and D.J. Wood, *REDEEM: Exploiting symbiosis between psychology and authoring environments*. International Journal of Artificial Intelligence in Education., 1997. **8**: p. 317-340.
6. Hsieh, P.Y., H.M. Half, and C.L. Redfield, *Four easy pieces: Development systems for knowledge-based generative instruction*. International Journal of Artificial Intelligence in Education, 1999. **10**: p. 1-45.
7. Luckin, R. and B. du Boulay, *Ecolab: The Development and Evaluation of a Vygotskian Design Framework*. International Journal of Artificial Intelligence in Education, 1999. **10**: p. 198-220.
8. Mark, M. and J.E. Greer, *The VCR tutor: Effective instruction for device operation*. The Journal of the Learning Sciences, 1995. **4**: p. 209-246.
9. Shute, V.J., *SMART evaluation: Cognitive diagnosis, mastery learning and remediation*, in *Proceedings of AI-ED 95*, 1995, AACE: Charlottesville, VA. p. 123-130.
10. Bloom, B.S., *The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring*. Educational Researcher, 1984. **13**: p. 4-16.
11. Cohen, P.A., J.A. Kulik, and C.C. Kulik, *Educational outcomes of tutoring: A metaanalysis of findings*. American Educational Research Journal, 1982. **19**: p. 237-248.
12. Graesser, A.C., et al., *Teaching Tactics and Dialog in AutoTutor*. International Journal of Artificial Intelligence in Education, 2001. **12**: p. 257-279.
13. Koedinger, K.R., et al., *Intelligent tutoring goes to school in the big city*. International Journal of Artificial Intelligence in Education, 1997. **8**: p. 30-43.
14. Wood, D., J. Bruner, and G. Ross, *The role of tutoring in problem solving*. Journal of Child Psychology and Psychiatry, 1976. **17**: p. 89-100.
15. Ainsworth, S.E., B. Williams, and D. Wood, *Using the REDEEM ITS authoring environment in naval training*, in *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, 2001. p. 189-192.
16. Ainsworth, S.E. and S.K. Grimshaw, *Are ITSs created with the REDEEM authoring tool more effective than "dumb" courseware?*, in *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, 2002, Springer-Verlag: Berlin. p. 883-892.