

Evaluating the REDEEM Authoring Tool: Can Teachers Create Effective Learning Environments?

Shaaron Ainsworth & Shirley Grimshaw, *School of Psychology, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom*
Shaaron.Ainsworth@nottingham.ac.uk
<http://www.psychology.nottingham.ac.uk/staff/Sharon.Ainsworth/>

Abstract. The REDEEM authoring environment allows teachers to create learning environments from existing computer-based training (CBT) by imposing their pedagogical preferences about how students should best be taught. We conducted two studies where classroom teachers constructed learning environments with REDEEM from pre-existing CBT to explore if this approach is educationally effective. Using a crossover design, the learning outcomes for 14-16 year old students who studied Genetics with these environments (either a REDEEM then CBT course or *vice versa*) were compared. In the first study, we found that performance of 74 students improved significantly but was not influenced by type of environment. Inspection of process data revealed that students who engaged with REDEEM's features did learn more. In the second study, conducted in a more natural context, a further 15 students completed the courses. REDEEM significantly improved learning compared to CBT. Analysis suggested that REDEEM enhanced performance by supporting additional interactivity but that macro-adaptation did not appear to impact upon learning. Possible interpretations of these results are discussed in the light of the many evaluation issues for authoring tools.

INTRODUCTION

Research has shown that when learners interact with Intelligent Tutoring System (ITSs), that such systems can generate impressive learning outcomes (*e.g.* Koedinger, Anderson, Hadley & Mark, 1997; Mark & Greer, 1995; Lesgold, Lajoie, Bunzo, & Eggan, 1992, Shute & Psotka, 1996). Despite this, ITSs are only just beginning to achieve application in schools, colleges and the workplace – perhaps because of the difficulty in developing them. It is estimated to take 200-1000 hours to create an hour of ITS instruction (*e.g.* Woolf & Cunningham, 1987; Murray, 1999). Consequently, one of the primary goals motivating the development of ITS Authoring Tools (ITSATs) is that of delivering the benefits of ITSs created in a cost and time effective manner.

Early ITSATs such as the Instructional Design Environment (Russell, Moran, & Jordan, 1988), KAFITS (Murray & Woolf, 1992) and COCA (Major, 1994) allowed teachers to construct appropriate domain material and to create their own teaching strategies. However, an evaluation of the authoring tools in COCA (Major, 1994) showed that despite offering considerable power to teachers, there remained a gap between the kinds of interfaces teachers would be prepared to use and the AI-based authoring tools that required them to express teaching decisions in pseudo-code. For example, a rule in COCA might be expressed as "IF summarized next concept AND taught next concept THEN next activity of session is test". A simple matter for AI researchers used to representing knowledge in production rules, but abstract and artificial for many teachers. REDEEM (**R**eusable **E**ducational **D**esign **E**nvironment and **E**ngineering **M**ethodology) was developed as a response to this evaluation. REDEEM reduces the teacher's opportunities to modify low-level instructional behaviour in favour of improving the ease of authoring so that classroom teachers have the opportunity to be seriously involved in the development of learning environments.

REDEEM represents one end of the continuum of the current generation of ITSATs (for a review of other approaches, see Murray, 1999, 2003). It does not support the construction of domain material, instead

focussing on the authoring of pedagogy. REDEEM learning environments have little domain knowledge compared to systems such as Demonstr8 (Blessing, 1997), which models the domain in terms of a detailed production system account of correct and incorrect behaviour or Diag with its knowledge of fault finding and diagnosis (Towne, 1997). They rarely include complex simulations of the sort that RIDES supports (Munro, *et al.*, 1997). The REDEEM tools are generic in terms of the domains to which they can be applied, so the benefits of knowledge rich tools have been sacrificed (Bell, 1998). Consequently, REDEEM creates learning environments that are adapted to the needs of students in ways that are impossible with conventional CBT. But, compared to traditional ITSs the learning environments are only minimally adaptive. In other words, REDEEM emphasises macro-adaptation (selecting an environment for a particular student) over micro-adaptation (a response to particular action such as selecting the next action be it a new question or type of hint).

To date, there have been few evaluations of the products of ITSATs to see if they deliver similar improvements in learners' knowledge and skills to ITSs created in more traditional ways. Probably the only ITSAT that has been substantially evaluated is XAIDA (Hsieh, Halff, & Redfield, 1998). Studies revealed that students could learn successfully with ITSs created with XAIDA but did not address the relative effectiveness of the ITSs compared to other forms of instruction such as CBT or classroom teaching. Determining the reasons for the success of an ITSAT is more complicated than ITS evaluation given the need to evaluate the authors' as well as the learners' experiences (*e.g.* Murray, 1997; Ainsworth, Grimshaw & Underwood, 1999). We have previously evaluated authors' experiences with REDEEM and shown that ITS production with REDEEM is relatively efficient. On average, only 90 minutes is required for training and the majority of REDEEM's tools are simple to use, especially those for macro-adaptation. Authoring with REDEEM is time efficient. Authors have taken between one and five hours to create an hour of REDEEM instruction from existing CBT. This is true for domains as diverse as "Shapes" in primary mathematics and "Communication and Information System Principles" (Ainsworth, Underwood & Grimshaw, 1999; Ainsworth, Williams & Wood, 2001). In all cases, these authors had no prior experience with developing CBT and in one case had rarely even used a word processor. We have now turned to evaluating learning outcomes with REDEEM learning environments.

Establishing causal relationships between aspects of the ITS design and positive learning gains is very difficult. To be done successfully, precise and large-scale experiments are often required (*e.g.* Shute, 1992; Mark & Greer, 1995). Furthermore, the effectiveness of any learning environment is influenced by the context of its use and so evaluations need to be conducted in real situations (*e.g.* Koedinger, *et al.*, 1997). For ITSATs, these methodological and philosophical issues are multiplied because an ITS is a combination of the options for authoring offered to users, the authors' decisions, and the systems' interpretation and delivery of those decisions. For REDEEM the problem is compounded still further as the resulting systems not only depend on the user, authoring tools and ITS shell but also involve externally imported domain material.

To try and determine the effectiveness of an ITS, experimenters have used a number of different alternatives as a control (for a review see, du Boulay, 2000). Many evaluations compare ITSs to classroom teaching (*e.g.* Shute & Glaser, 1990; Meyer, Miller, Steuck, & Kretschmer, 1999). Famously, Bloom (1984) 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 0.4 sigmas (Cohen, Kulik, & Kulik, 1982). Evaluations of ITSs reveal effect sizes of between 0.4 and 1 compared to classroom teaching (*e.g.* Graesser, Person, Harter, & Tutoring Research Group, 2001; Koedinger, *et al.*, 1997). We could realistically hope that REDEEM learning environments will fall around the 0.5 area rather than 1.0, as they are not particularly sophisticated ITSs.

Another common technique is to use within system comparisons (*e.g.* Ainsworth, Bibby & Wood, 2002; Corbett & Anderson, 1991; Arroyo, Beck, Woolf, Beal, & Schultz, 2000). This allows investigation of the contribution of different design decisions on learning outcomes and whether there are Aptitude-Treatment interactions (ATI) where one type of learner benefits from one version of the environment and another type of learner, an alternative version. For example, Shute (1993) contrasted two versions of the Flight Engineering Tutor, which varied the number of problems the tutor required students to complete. Overall, there were no differences in learning outcomes associated with the type of tutor, but there were ATIs (*e.g.*

High Knowledge, Low Working Memory students learnt better with few problems). Within system comparisons are particularly suitable for exploring if there are additional benefits to macro-adapting the style of a learning environment to an individual learner's needs.

A variation on this technique is ablation experiments where particular design features are removed and performance of the resulting systems compared (*e.g.* Cohen, & Howe, 1988). They also allow analysis of the contribution that specific system features bring to the learning experience. For example, Mark and Greer (1995) compared four versions of a tutor that taught learners how to operate a video recorder. The "smartest" one used model tracing to monitor learners' performance and could give detailed feedback on misconceptions. The "dumb" version allowed learners only one way to perform a task and provided only simple prompting. The smarter systems decreased the number of steps, errors and time required for students to complete the post-test. Shute (1995) assessed the contribution made by the remediation component of Stat Lady, which detects if learners are not succeeding at a curriculum element and adjusts instruction accordingly. Shute found that by enabling remediation, learning times increased but so did learning outcomes. Luckin & du Boulay (1999) examined three versions of ECOLAB, which varied in how much responsibility the system took for deciding on such factors as the help offered and the abstraction of terms. They found a complex pattern of results but showed overall that learners made more productive use of the system and generally learnt more with the "intelligent" versions. Overall, these studies have tended to show that systems with more "intelligence" create more effective learning experiences for students.

In this paper, we report two studies where secondary (high) schoolteachers were given two previously developed courses that teach the age 14-16 UK curriculum on the topic of Genetics. They were asked to author their ideal ITSs for their students who subsequently took part in learning outcome studies. As REDEEM is based on non-intelligent CBT, we can essentially perform a massive ablation and produce both a "dumb" and "smart" version of the same course. Then the learning outcomes from those students working with CBT can be compared to learning outcomes with REDEEM learning environments. If learning outcomes are higher with REDEEM, then the conclusion that the REDEEM/Author partnership in the situation provided better support for learning than the non-intelligent courseware is warranted. If authors create different ITSs for learner groups, then it may also be possible to examine what aspects of teaching strategies are beneficial for (particular groups of) learners. Thus, these studies involved both ablation (stand-alone courseware versus REDEEM learning environments) and within-system comparisons (multiple REDEEMs with differing teaching strategies). Study one was conducted under experimental conditions and Study two was performed in a school classroom. Before describing these experiments, we present a brief overview of how REDEEM works.

SYSTEM DESCRIPTION

The REDEEM suite was developed in Click2Learn ToolBook Instructor and runs on Windows 95+. It consists of three main pieces of software - courseware catalogues, authoring tools and ITS shell (Fig. 1). Authors use the REDEEM tools to describe courses, supplement them with learning activities, construct teaching strategies, and identify types of students. The REDEEM ITS shell uses this knowledge, together with its own default teaching knowledge, to interpret the courseware in such a way as to deliver adaptive, interactive instruction. The shell's role is to sequence this material for different users, provide a number of teaching strategies and additional questions with feedback, monitor student performance, macro-adapting the environment if requested, support integration into classroom teaching by the use of non-computer based tasks and reflection points and provide teachers with detailed feedback on students' performance. We will briefly describe the main components, but for a fuller description, the reader is referred to Major, Ainsworth & Wood (1997) or Ainsworth *et al* (2003).

Courseware catalogues

Domain material in REDEEM is based on the idea of a courseware catalogue. It consists of pages from CBT developed in either Click2Learn ToolBook or in HTML. Consequently, this severely limits the flexibility of

the resulting ITS. However, it does allow greater reusability, and, of course, significantly reduces the time to create an ITS compared to creating the domain material from scratch. The ideal courseware for REDEEM presents discrete pages of material showing different aspects of the domain at varying levels of difficulty. Typically, REDEEM has been used to present primarily declarative material, as although pages can contain multi-media, simulations, animations, questions and exercises, REDEEM does not model the learners' actions on these objects. Before authoring with REDEEM begins, the authors identify the courseware they intend to import into REDEEM – for example by placing all the web pages into a single directory.

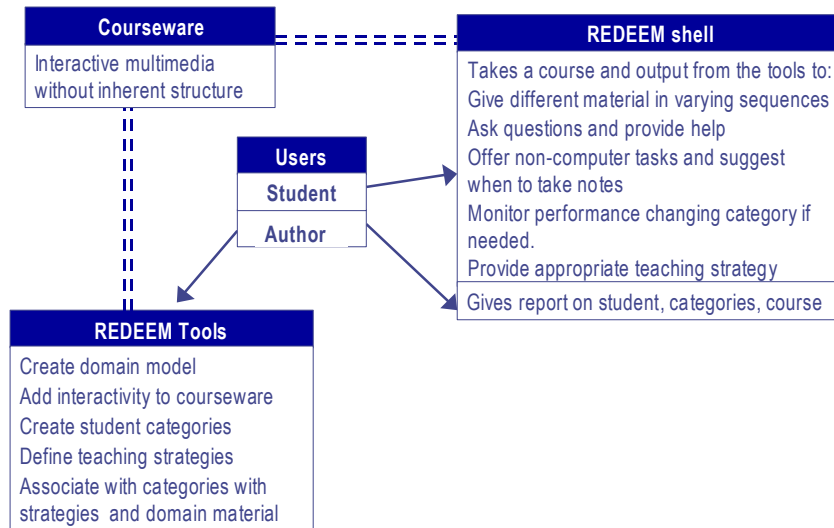


Fig. 1. REDEEM schematic

Authoring tools

REDEEM's authoring tools decompose the teaching process into a number of separate components. Essentially authors are asked to describe what they are teaching, whom they are teaching and how they would like to teach these students. This information is then combined by assigning particular teaching strategies and types of material to different student groups.

What to teach

One of the most important stages in authoring involves the description of the course material. The first task is to give each page a learner appropriate name; other tasks can then be performed in an order that authors wish. Pages are combined into sections. They can be placed in multiple sections, which need not consist of contiguous pages in the underlying CBT. Sections are then described upon a number of dimensional ratings, (*i.e.* they describe how familiar, easy, general or introductory a section is likely to be to their students). This is done by graphical manipulation of sliders. In addition, authors may describe relations between sections. The most commonly used relation is the prerequisite relation, which ensures that a section is not offered to students until prerequisite sections have been completed. Pages themselves are then described in terms of the same dimensional ratings and relations. Normally, these descriptions are considered in relation to other pages

in the same section rather than in comparison to pages across the whole course. Accordingly, a relatively complex page in an easier section would still be marked as difficult. Relations between pages are only supported within a section. These tools provide information that the system uses as a semantic network describing the structure (rather than the content) of the teaching material. This network expresses syntactic information about the properties of page and section contents and relations between these pages and sections. There are three levels to this network, which represents a compromise between additional flexibility and ease of authoring. This network enables the shell to make default decisions about adapting content and to implement teachers' preferred routes through material.

The next stage is to add interactivity. Authors can associate a reflection point or non-computer task with a page. They can create questions (multiple choice, fill in the blank, multiple true, true-false or matching questions). For example a typical question was "Cats are members of the same species because they" and up to five answers such as "are the same size", one of which should be the correct answer "are able to breed and produce fertile offspring". They provide feedback which will explain to the student why that answer is correct "Well done - scientists define a species as a group of living things that can breed and produce fertile offspring". In addition, an important aspect of the REDEEM approach is the ability to offer learners multiple levels of help in way that is similar to contingent help (Wood, Bruner & Ross, 1972). The author can create up to five different hints for each question, ranging from very general hints such as "think about the differences between Siamese and Burmese cats" to very specific hints such as "When cats breed they produce kittens which can produce kittens of their own". Authors describe a number of characteristics of the question that the ITS shell uses to implement a specific teaching strategy. They assign a difficulty level to the question, decide whether it should be offered before or after the page (pre-test or post-test) and state whether its position is constant or can vary with the teaching strategy.

Who to teach

Students are described as belonging to one of a set of author-defined categories. Categories can be at any degree of granularity, ranging from a whole class to an individual student. Commonly, teachers have tended to use performance-based measures (*e.g.* high flyer, struggler) or task-based measures (*e.g.* revising) or have combined these (*e.g.* high reviser). However, it is possible to use any dimension that authors find appropriate. If the author wishes, the validity of (performance-based) categories can be evaluated against students' question performance. In this case, the shell will automatically change the category as the overall standard of the student (as defined in the shell's student model) changes. This can result in a new teaching strategy.

How to teach

The third important aspect of the authoring task is the definition of a number of teaching strategies as the basic repertoire of the ITS shell (see Fig. 2).

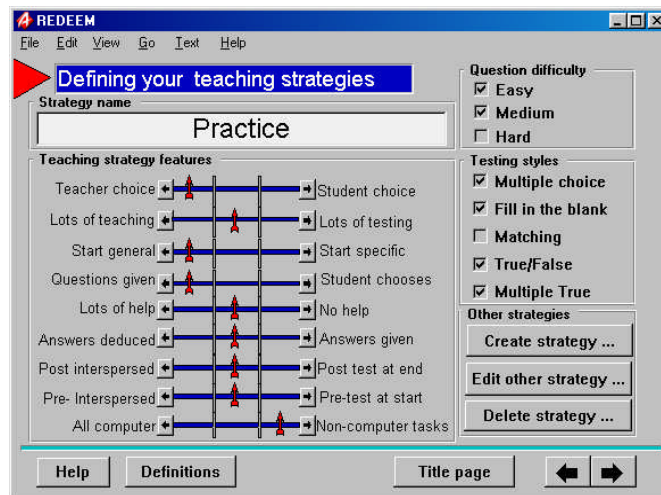


Fig. 2. Creating a teaching strategy

Different instructional principles can be embodied in various strategies by manipulating the sliders. Each slider in Fig. 2 has three discrete positions that result in different instruction. Consequently teachers are free to create as many strategies as can be composed from the various instructional attributes. Questions are also associated with teaching strategies in terms of their difficulty (*e.g.* include easy and medium questions) or their type (exclude matching questions). For example, teachers may choose to create a “Free Discovery” strategy where student chose the material they see, the order they see it, the questions they answer, the number of attempts they have at each question, when to receive help and whether to perform non-computer-based tasks. In contrast, the teacher may chose to create a “Guided” strategy where students have no choice over material, when questions of only certain types and difficulties are included and asked immediately after the relevant material has been presented, and help is given on error with a limited number of attempts for each question. Teachers can create as many strategies as they wish. In fact, REDEEM can offer 10000 different alternatives each subtly different to each other, although to date no author has created more than seven (see Ainsworth *et al* (2000) for a study which analysed teachers’ choice of strategy).

What students learn

Authors differentiate material for learners by associating sections with student groups. By default, learners see all the material, but the author can choose to remove sections for a particular student category (*e.g.* to focus on introductory material for those who need more help).

How students learn

The final necessary stage of authoring is to relate each student category to a teaching strategy. To date, authors have varied from creating a single preferred strategy to creating a unique strategy for each group. They have based their decisions on the perceived knowledge or abilities of students.

ITS shell

The ITS shell delivers the courseware according to the output of the authoring tools in combination with its predetermined defaults.

Delivering adapted and/or adaptive instruction

The main role for the shell is to deliver the course material to each student in the way that the teacher specified with the authoring tools (Fig. 3). Tutorial actions available to the shell (depending upon the teaching strategy) are: to teach new material; offer a question (and help if appropriate); suggest that students make notes with the on-line tool; offer a non-computer based task and by means of password protection check that it has been completed; or summarize students' progress. The two most complex actions are teaching and questioning.

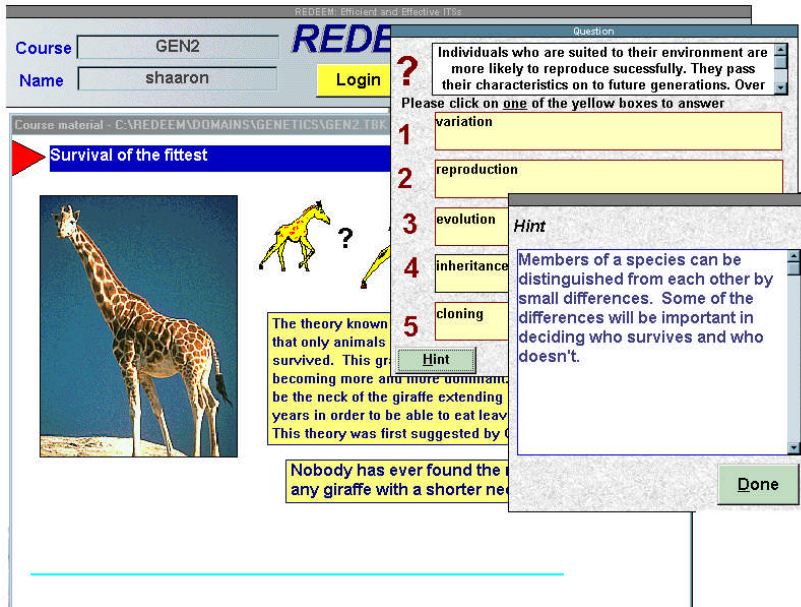


Fig. 3. REDEEM shell running genetics2

If the shell is teaching, it computes a weighted array of choices using the semantic network of pages and its default assumptions (e.g. prefer easy before difficult material or familiar before unfamiliar). This is done both at the section and page level. Other rules check that pages in the same section are together and that prerequisite pages are taught in the correct order. The way that this weighted array is used depends on the level of student control. If the student control is set to high, the student is presented with the most appropriate page. If set to partial student control, the learner chooses sections but has no choice within a section. If set to full student control, then learners are presented with a hierarchically presented menu of the complete course organized according to the weighted array. Questions are selected and offered to learners in a way that depends on many interacting factors of the teaching strategy, i.e. question choice, amount of questions, pre-test position, post-test position, level of difficulty and type of appropriate question. The support the ITS shell offers to students for answering questions also depends on the strategy. It must compute the number of attempts per questions a learner is allowed and whether to offer help on request or on error.

Student modelling and history

REDEEM employs a basic overlay model that records the system's understanding of the students' knowledge of an area. The values of the model change over the course of a session as the student sees new material and answers questions. This model is primarily used when student categories are performance based

and determines if learners should change student category. The shell also maintains a student history. This is used to offer reports to the author either on an individual student's progress, a student category's progress or to give a report on the course. To do this, the shell keeps a trace of all modules taken, including pages visited, questions that were asked and their answers, number of hints offered, scores and time on tasks. Teachers can use this information to monitor the progress of learners, for example to see if they require multiple attempts to get questions right, use help appropriately, *etc.* Student category reports allow teachers to compare the performance of a group of students (*e.g.* to determine if one student is falling behind). The course report allows teachers to see an overall picture of how their class is progressing and is particularly useful for examining the way that particular questions are answered to assess if they are at an appropriate level of difficulty.

STUDY 1

Authoring phase

A science teacher was recruited to update a previously designed genetics course so that it met the requirements of the UK's National Curriculum and the GCSE syllabus. The course consisted of declarative material with some multimedia, simple exercises and a glossary. New material was added and old material removed if considered no longer relevant to the syllabus. The course was divided into two parts. The first, Genetics1, was 48 pages long and covered the topics of inheritance, genes, and cell division. The second, Genetics2, was longer at 73 pages, and covered DNA structure, evolution and reproduction.

Table 1
Learning environments created for five student categories for Genetics1 & Genetics2

	Group A (N= 8)	Group B (N= 30)	Group C (N= 23)	Group D (N= 15)	Group E (N= 10)
<u>Content</u>					
<u>Difficulty</u>	difficult	quite difficult	easier	easier	easier
<u>Amount</u>	44 & 60 pages	44 & 50 pages	32 & 44 pages	30 & 44 pages	30 & 44 pages
<u>Questions (Qs)</u>					
<u>Types</u>	all types	all types	all types	no matching	no matching
<u>Difficulty</u>	med. & hard	med. & hard	easy & med.	easy & med.	easy & med.
<u>Amount</u>	36 & 39 Qs	36 & 39 Qs	24 & 24Qs	23 & 24 Qs	23 & 24 Qs
<u>Limit</u>	all	all	1 per page	1 per page	1 per page
<u>Strategy</u>					
<u>Content</u>	choose sections	choose sections	no choice	no choice	no choice
<u>Question</u>	selects Q type	selects Qs	Q after section	Q after section	Q after page
<u>Help</u>	on error	on error	on error	error & request	error & request
<u>Ans-deducted</u>	many tries at Q	many tries at Q	many tries at Q	2 tries at Q	2 tries at Q

The teacher then used the REDEEM tools to create learning environments for the two Genetics courses. The material was described to create the semantic network for the REDEEM shell. Sections were developed, often addressing the same topics but at varying levels of complexity. The teacher created 69 questions and multiple levels of hints to their solution. She identified reflection points and developed a number of non-computer tasks (*i.e.* worksheets), which were authored to appear at appropriate times. She then chose how to

adapt the learning environments to the perceived needs of different students. She created five different categories of students based on her judgments of their relative aptitude (A to E). Each category was assigned different content and teaching strategies (see Table 1). Some aspects of her strategy such as using no pre-tests, assigning non-computer based tasks and teaching from general to specific concepts were common to all strategies. These tasks took twelve 90 minutes sessions.

CBT courses

Two CBT courses were constructed from the underlying courseware to contain material that the teacher felt was essential for all ability groups. Out of a possible 48 pages from Genetics1, 33 were included in CBT Genetics1 and 44 from 73 were included in CBT Genetics2. The order of presentation of the pages was fixed by the teacher and navigation was limited to “go next page” and a hotlink to glossary. A workbook was created to contain the same non-computer tasks as those in REDEEM and some blank pages to make notes. Hence, the CBT courses were similar to the REDEEM courses but lacked REDEEM’s interactivity and macro-adaptation.

METHOD

Design

In order to reduce the effects of participant variance, a crossover design was employed. All participants received one course under REDEEM and one as CBT, *i.e.* half received REDEEM Genetics1 and CBT Genetics2 and half CBT Genetics1 and REDEEM Genetics2 (see Fig. 4). Subjects’ scores at pre-test were used to ensure that there was an equal distribution of ability across the two conditions.

Fig. 4. Design of study

Participants

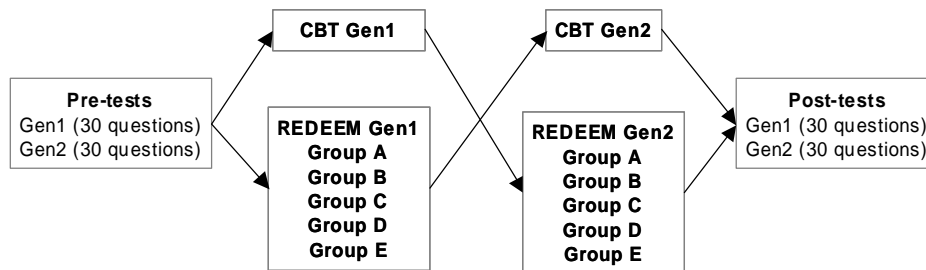
Eighty-six students from a mixed ability state secondary

school took part in the experiment. They were 14 and 15 years old and there were 45 boys and 41 girls. No attempt was made to have equal numbers assigned to different categories, as this was an important part of the author’s teaching strategy. The number in each category was 8 (A), 30 (B), 23 (C), 15 (D) and 10 (E).

Materials

Developing comparable pre and post-test material is particularly complex given the differentiation by content and strategy that REDEEM provides. Given the lack of domain taxonomy for Genetics, we decided the best solution was to use questions on the pre and post-tests that were based on material presented to all subjects and that were repeated on the pre and post-test. Whilst it might be possible in problem solving and procedural domains to identify parallel items, this is not the case with declarative topics where you either know a fact or you do not. Furthermore, some of material is directly questioned for (some) groups by REDEEM. Rather than excluding the material, we chose to address this issue by creating three types of question:

- REDEEM questions - all participants were asked these during their REDEEM sessions;



- Near Transfer questions – which addressed the same issue as a REDEEM question but were manipulated so that memorization was not sufficient (e.g. the inheritance of brown or blue eyes was mapped onto aliens with purple and pink eyes);
- Non-REDEEM questions – the material to answer the question was presented to all students but never directly questioned.

In total, a 60 item multiple choice quiz was developed (one correct answer and three distracter items). It consisted of 30 questions on Genetics1 and 30 on Genetics2 each further subdivided into 10 REDEEM, 10 Near Transfer and 10 Non-REDEEM questions. There were two versions of the quiz such that half the participants answered questions on Genetics1 followed by questions on Genetics2, and *vice versa*.

Procedure

1. Pre-tests were given to the participants in their school classroom just prior to the intervention.
2. Intervention: The students came to the University of Nottingham to study the Genetics material. Each session lasted between 30 and 90 minutes depending on whether students were available for a single or double lesson. The minimum number of sessions a student attended for was three and the maximum was five. Two different computing labs were used each equipped with up to 32 PCs. There were up to three experimenters and two teachers on hand to deliver non-computer tasks, provide help with the interface to the software and provide classroom management. Participants were provided with instruction booklets to help them navigate through the courses. No direct teaching of the concepts took place.
3. Post-tests were given to the participants within two weeks of their finishing the study.

RESULTS

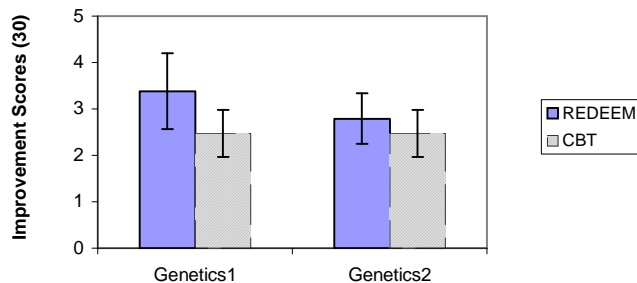
Learning outcomes

To examine the effects of the intervention, a [2 by 2 by 2] ANOVA was carried out on the pre-test and post-test data. The design of the analysis was 2(Genetics1, Genetics2) by 2(pre-test, post-test) with a between-subjects factor of order (REDEEMGenetics1/CBTGenetics2, REDEEMGenetics2/ CBTGenetics1). Twelve subjects were excluded due to non-completion of a test (Table 2).

Table 2
Pre and post test scores (out of 30) by course and type of environment

	REDEEM		Genetics2		CBT		Genetics2	
	(n = 40)		(n = 34)		(n = 34)		(n = 40)	
	\bar{x}	S.D.	\bar{x}	S.D.	\bar{x}	S.D.	\bar{x}	S.D.
Pre-test	11.00	3.37	12.68	3.94	11.71	3.01	11.83	3.69
Post-test	14.38	4.83	15.47	4.63	14.18	3.75	14.30	3.69

There was a significant main effect of time ($F_{1,72} = 74.52$, $MSE = 7.62$, $p < 0.001$), that is, as predicted, post-test totals were greater than pre-test totals (on average scores rose from 40% to 49%). There was also a significant main effect of course ($F_{1,72} = 5.06$, $MSE = 8.25$, $p < 0.05$) with students scoring higher on



Genetics2 than on Genetics1. There were no significant interactions. Fig. 5 graphs this as improvement scores (e.g. scores on Genetics1 were 3.4 higher on the post-test than they were on the pre-test for the REDEEMGEN1 condition) and with error bars to show standard error of the mean. There was no correlation between students' improvement on Genetics1 and Genetics2 ($r = 0.12$) and so students who made the greatest improvement on their REDEEM course were not the same ones who made greatest improvement on their CBT course.

Fig. 5. Improvement scores by environment and course.

An analysis of the effect of question type (REDEEM, Near Transfer (NT) and Non-REDEEM (Non)) was performed. We predicted that for the questions on the course they had experienced through REDEEM, students would perform significantly better on REDEEM and NT questions than on Non-REDEEM questions. There should be no difference between questions types for CBT material as no questions were asked. Two [2 by 3 by 2] ANOVAs were performed on the REDEEM and CBT data respectively, with two within-subjects factors, time and question type and one between-subjects factor, course. For the REDEEM data, there was a significant main effect of time ($F_{1,72} = 41.25$, $MSE = 2.83$, $p < 0.0005$) and question type ($F_{2,144} = 9.99$, $MSE = 2.45$, $p < 0.001$). There was a significant interaction between question type and course ($F_{2,72} = 4.79$, $MSE = 2.78$, $p < 0.02$) (see Fig. 6 where the interaction is graphed as improvement scores). Simple main effects showed that question type had a significant impact only on Genetics2 ($F_{2,144} = 12.40$, $p < 0.001$). Furthermore, there was a significant interaction between time and question type ($F_{2,144} = 5.11$, $MSE = 1.67$, $p < 0.01$). Simple main effects analysis revealed that there were no differences between the question types at pre-test ($F_{2,288} = 1.28$) but that there were at post-test ($F_{2,288} = 13.64$, $p < 0.001$). Post hoc comparisons revealed that this was because REDEEM and Near Transfer questions showed significant improvement from pre-test to post-test ($q = 7.54$, $p < 0.001$ and $q = 5.67$, $p < 0.001$) whereas non-REDEEM questions did not ($q = 2.70$). Analysis of the CBT data showed a single significant effect, that of time ($F_{1,72} = 43.71$, $MSE = 1.71$, $p < 0.001$)

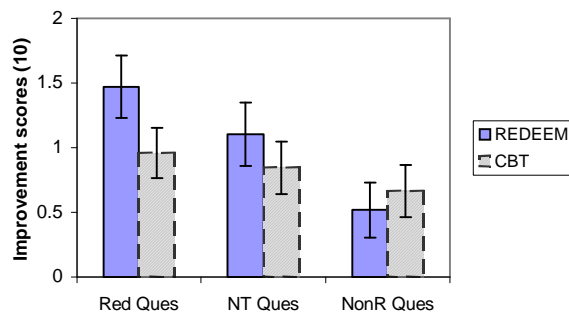


Fig. 6. Improvement scores by question type and environment (collapsed across course).

Relation between pre-test scores, teachers' categorisation and learning outcomes

The relation between pre-test scores and learning outcome was examined. There was a significant positive correlation between pre-test and post-test scores ($r = 0.70$, $N = 74$, $p < 0.001$), but no significant relationship between pre-test scores and improvement scores, ($r = -0.10$); students at all levels of prior knowledge made similar improvements from pre to post-test.

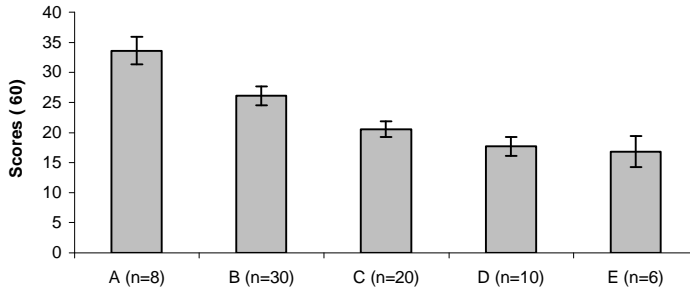


Fig. 7. Pre-test scores (out of 60) by student category

Fig. 7 shows the scores for the different teachers categories based again on their performance on the pre-tests. Students' scores were related to the category such that the scores were Group A > Group B > Group C > Group D > Group E (Jonckheere-Terpstra = 140.5, $p < 0.001$). Tukey tests showed that the two higher groups differed to each other and to the three lower groups (*i.e.* Group A with B,C,D,E and Group B with C,D,E; $p < 0.001$ in all cases).

To determine if any student category had differentially improved, pre and post-test performance was examined. To achieve a sufficiently large number of subjects per cell, the cells were combined to create two ability-groups of high and low scorers (A/B and C/D/E). The pre-test scores of Groups C,D and E do not differ from one another, but they do differ to groups A and B. These groups also receive similar teaching strategies to each other (see Table 1). As previous analyses had confirmed no interactions between course and time, we collapsed across course. Consequently, the design for the analysis was 2(REDEEM, CBT course) by 2(pre-test, post-test) with a between-subjects factor of ability-group (High, Low).

Analysis by [2 by 2 by 2] ANOVAs on the subjects' REDEEM and CBT test data showed a significant main effect of time ($F_{1,72} = 75.45$, $MSE = 7.64$, $p < 0.0001$) and ability -group ($F_{1,72} = 63.76$, $MSE = 21.02$, $p < 0.0001$). None of the interactions were significant. Both ability groups improved similarly when learning with REDEEM and CBT. This is graphed as improvement scores in Fig. 8.

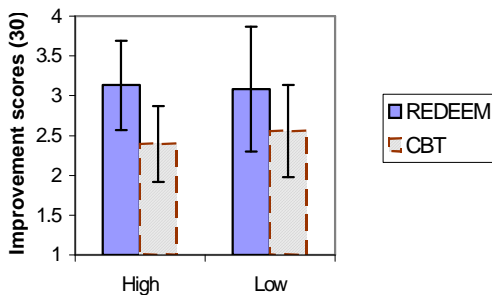


Fig. 8. Improvement scores by ability-group and by environment (collapsed across course)

Process measures for REDEEM and CBT

Some students in this study learnt a considerable amount (the greatest improver increased their score by 17) and some failed to learn or got worse (the worst improver decreased their score by 8). Consequently, we need to determine how the students' use of the software influenced learning. To address this question a number of different interaction measures were explored.

To calculate the amount of time the participants had spent learning, the time they had spent each session from their first to their last mouse click was totalled. The time per course is influenced by the fact that each course differed in length and the REDEEM courses also differed depending upon the student category. Hence, total time was divided by the number of pages in the course to determine the mean amount of time per page. Analysis by [2 by 5 by 2] ANOVA showed a significant main effect of group ($F_{1,61} = 2.95$, $MSE = 3197.898$, $p < 0.05$). Post-hoc tests revealed that students in Category A spent significantly longer per page than students in Groups D and E ($q = 4.02$, $p < 0.05$ and $q = 4.71$, $p < 0.05$)(Fig. 9). There was a significant interaction between environment and course ($F_{1,67} = 8.694$, $MSE = 3197.90$, $p < 0.005$). Simple main effects analysis confirmed that students spent longer learning with REDEEM on Genetics1 ($F_{1,22} = 6.36$, $MSE = 3515.98$, $p < 0.02$) but that there were no differences on Genetics2.

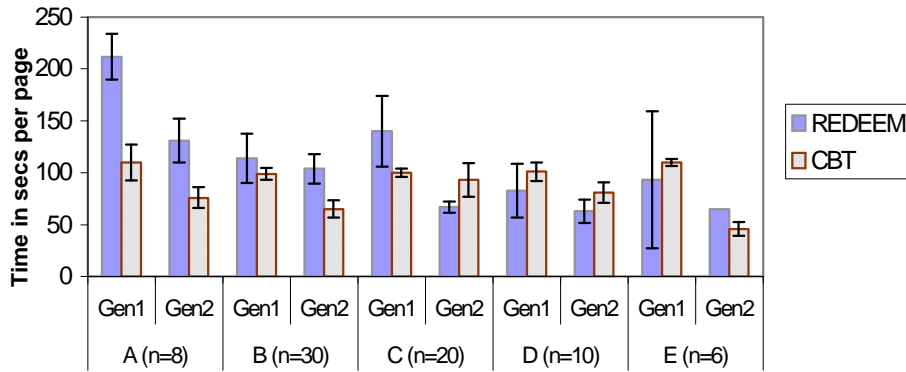


Fig. 9. Time per page by Environment, Course and Student Category

We inspected this data to see if students who spent longer on the course showed differential improvement. There was no systematic relationship between performance (pre-test, post-test, improvement) and time for the CBT groups, but there was for the REDEEM groups; increased time was positively associated with improvement ($r = 0.29$, $N = 74$, $p < 0.02$) and post-test performance ($r = 0.33$, $N = 74$, $p < 0.005$).

Use of notes

Students were provided with a pen and paper when learning with the CBT and an on-line notes tool in REDEEM. They were told that writing notes would help them understand and remember the material. To test the validity of this statement, we performed a simple analysis of students' notes exploring the amount of notes written rather than the quality of those notes.

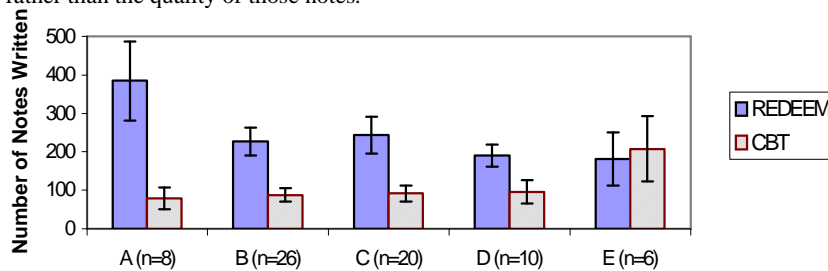


Fig. 10. Number of words written by student category and environment

NB Four students in Group B removed their notebooks in the CBT condition and have been excluded from the analysis.

A [5 by 2] ANOVA analysed if there was differences in the number of notes written by student category and environment. Students wrote significantly more when learning with REDEEM then when learning with CBT ($F_{1,64} = 17.65$, $MSE = 26246$, $p < 0.001$). There were no further effects. Pearson's correlations were carried out on the number of notes and performance (pre-test, post-test, improvement). There was a weak but significant relationship between note taking and learning outcomes for the REDEEM group. Students who wrote more notes performed better on the post-test in the REDEEM condition ($r = 0.30$, $N = 72$, $p < 0.01$) even when pre-test scores were partialled out ($r = 0.22$, $N = 72$, $p < 0.05$). However, there was no relationship between performance and note taking in the CBT condition.

REDEEM only process measures

REDEEM automatically captures information about an individual's use of system features as these form the basis of student reports. In addition to the information about time spent reading information or responding to questions, it also automatically logs the number of attempts at questions, questions scores, and the number of hints either provided or requested. Analysis of these measures allows us to explore how these system features were used and whether there was a systematic relationship between behaviour with REDEEM and performance. These analyses are presented by student category and have for simplicity been collapsed across course.

Question answering

To explore if there was a relationship between the students' performance on questions during their intervention session and their incoming knowledge or post-intervention performance, we analysed their responses to questions. Questions were classified into those right first time and everything else which could include answers right on a second, third attempt or no correct answer.

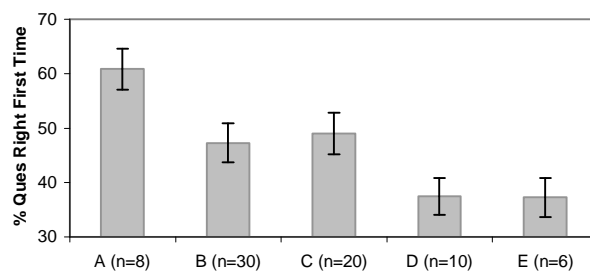


Fig. 11. Percentage of questions right first time by student category

A [5 by 1] ANOVA on percentage of questions right first time confirmed the influence of category ($F_{4,67} = 4.31$, $MSE = 179.7$, $p < 0.005$) with students in Category A performing better than those in D or E ($q = 23.41$ $p < 0.004$ and $p = 23.61$, $p < 0.02$). There was a consistent relationship between question behaviour and students' prior knowledge. Those students who scored higher on the pre-test were more likely to answer the question right first time ($r = 0.33$, $N = 72$, $p < 0.001$). There were also significant correlations remain between post-test and improvement scores, and percentage of questions right first time, ($r = 0.63$, $N = 72$, $p < 0.001$ & $r = 0.43$, $N = 72$, $p < 0.001$). If pre-test scores are partialled out, the significant relationship between answering questions correctly and improvement scores remains ($r = 0.57$, $N = 72$, $p < 0.001$).

Comment [sea1]: Or 67?

Use of help

The teacher had authored a number of hints for questions in REDEEM. The way these hints are made available to students depends on two dimensions of the teaching strategy. The major determiner is “amount of help” and for groups A to C, the teacher chose to use hint on error, so REDEEM only shows a hint when a learner gets the answer wrong. In these cases, amount of help is essentially the same as number of attempts at question and so is not analysed further. In contrast, groups D & E are allowed to request help.

Table 3
Number of hints requested by student category

	No of Qs	Total no of hints on request		% Qs where hints requested	
		\bar{x}	SD	\bar{x}	SD
D (n= 10)	23 or 25	4.6	6.88	11.62	12.23
E (n= 6)	23 or 25	3.0	3.46	9.51	12.00

Table 3 shows that there was a low uptake of hints on request. This is not because students knew the answers to these questions as Fig. 11 reveals that students in these categories were only likely to get the question right first time around a third of the time. There was also a large standard deviation with most students requesting only a very few hints (four or less over the whole course) and a couple of students requesting over 20. There was no significant correlation between hints on request and any measure of performance.

DISCUSSION

Learning outcomes

The results of this study showed that whilst students in all conditions improved their knowledge of genetics, there was no differential impact of REDEEM on learning outcomes. Students’ pre-test to post-test improvement was the same whether they received the course as CBT or as REDEEM. This was true for learners in all student categories. Furthermore, the degree of pre-test to post-test improvement was statistically significant but not as substantial as we would have liked. Students’ scores for the material they learnt with REDEEM were an average of 3.11 questions better at post-test and for CBT material, the improvement was 2.47 questions from 30. When the pen and paper tests were examined more closely, it was evident that the students improved more on questions (and near transformations of questions) that REDEEM had presented during their intervention. We also examined students’ performance according to their student category. It was evident that the teacher had good knowledge of the likely performance of her students as the relationship between pre-test scores and student category was significant. However, there was no evidence of differential improvement for the student categories; both high and low performers learned the same amount.

REDEEM and CBT process measures

Typically time on task predicts learning outcomes and we inspected the data to see if this was true. Time on task (as measured by time per page) did not correlate with improvement for students learning with CBT, but it did for students with REDEEM. REDEEM supports more learning activities than the CBT (answering questions, reading feedback, prompts to write notes, etc). Unsurprisingly therefore students learning with REDEEM spent an average of 17 more seconds per page than students interacting with CBT. However, the significant correlation between time and learning outcomes in only the REDEEM condition suggests that its not time on course *per se* that is an important determiner of learning outcomes, but it is the time spent in active learning that is important.

To gain some insight into how students’ note taking influenced learning an analysis of their notebooks was performed by counting the number of words written without examining either the accuracy or quality of

the statements. Students made many more notes using REDEEM than they did when learning with the CBT. In addition, there was a significant relation between the amount of notes written with REDEEM and post-test performance but no relation in the CBT condition. Overall, REDEEM encourages students to write notes, and furthermore writing notes is a reasonable predictor of learning outcomes. One explanation that is consistent with this result is that REDEEM provides students with scaffolding about the most important concepts to write notes about and those students who responded to these prompts wrote more appropriate notes. Taken together with the time analysis, it would appear that if students take the opportunities that REDEEM affords them to interact more deeply with the material this can significantly enhance their learning outcomes. However, students who don't write notes and who don't engage with the material may not perform any better than when learning with the CBT.

REDEEM only process measures

A number of process measures that describe students' interactions with REDEEM were analysed. There was a significant positive relationship between the number of attempts needed to get a question correct with post-test scores that remained even when pre-test scores were partialled out. It would seem that students who needed fewer attempts to respond with the correct answer during the REDEEM sessions, learnt more from the experience than those who needed multiple attempts (in other words errors are associated with lower performance). It should be noted that REDEEM does not allow students to proceed past a question without indicating the correct answer(s) and explaining why that answer is right. Hence, all students should have had equal opportunity to learn from answering questions whether they get the answer right or wrong. However, this does depend on them reading the feedback messages. Students in Category A got more of their answer correct first time than those in the lower groups (significantly so with D & E). It would appear that questions were differentially easier for these students even though they were given medium and hard questions. However, it would be hard to argue that questions were too easy as performance was still at only 60% right first time.

Students could also receive help when answering questions and category D & E students were able to ask for help on request. They rarely took advantage of this feature. Students requested hints on only 10% of questions even though their first answer to questions was wrong around 62% of the time. Given this low uptake and the skewed distribution (one student accounted for 35% of all requested hints), unsurprisingly there was no relationship with performance measures. Students may not have found the hints useful, may have been unconcerned about their performance, may not have realised that they needed help, or preferred to try and find out the answer for themselves. Nonetheless, the limited use of this feature is interesting given that the teacher did not include it for all student categories as she was afraid of students choosing to ask for help rather than attempt the question. This fear does not seem to have been warranted.

In summary, analysis of the REDEEM and CBT process data indicate that certain students were more likely to improve their performance than others and these students were the ones who took advantage of REDEEM's features. Students using REDEEM who took more notes, spent longer learning and answered questions correctly were more likely to learn than those who did not. Whilst this was often related to pre-test, as those students who scored higher at pre-test tended also to engage in this behaviour, it was not just determined by prior knowledge as when pre-test scores were partialled out many of these factors remain significant.

Overall, the results of the study did not find that REDEEM learning environments were any more effective than the CBT that they were based on nor were any significant differences found between learning outcomes for the different REDEEM learning environments. One plausible explanation for this result that we sought to rule out was whether the unnatural situation of using university laboratories instead of school classrooms had reduced the validity of the study. We had been unable to use school computing facilities because of timetable clashes but bringing students into the University was not ideal. Firstly, given the time constraints this imposed, a single intervention session lasted up to 90 minutes and in some cases we held multiple classes on one day. Secondly, students found learning with the software as being somewhat removed from their everyday schooling. They felt somewhat disappointed that an exciting trip out of school led only to learning at a computer. It also meant that they viewed this experience as adjunct to their required

schooling and in some cases, this led to a significant lowering of motivation. For these reasons, we decided to repeat the study, but this time in a school classroom.

STUDY TWO

Authoring phase

The basic material was the same as Study One. However, this school followed a different syllabus and so the courseware was modified to take this into account. A class teacher recruited from the participating school then used the REDEEM tools to create his learning environments. He was provided with the learning environments from the Study One and maintained many of the features. However, he substantially changed the structure of the material (see Ainsworth, Clarke & Gaizauskas, 2002) and rewrote some of the hints. This teacher chose a coarser-grained description of students than in Study One. Three different categories of students were created that corresponded to different sets (classes differentiated by ability) at the school. These are labelled 1 to 3 to avoid the implication that there is any correspondence to the categories in Study One. In keeping with this less differentiated approach, the learning environments created with REDEEM were more similar to one another with only limited different material and questions. Furthermore, the teacher chose to give all groups the same teaching strategy (summarized in Table 4). This teacher chose to do much of his planning away from REDEEM and brought in paper notes. As a result, we can not quantify the time spent creating the learning environments. However, it was substantially less than the first teacher, partly because he reused many of the features.

Table 4
Learning environments created for three student categories for genetics1 & genetics2

	Group 1	Group 2	Group 3
Content			
Difficulty	most difficult	medium	easiest
Amount	44 & 53 pages	39 & 51 pages	34 & 48 pages
Questions (Qs)			
Types	all types	no multi-true	no multi-true
Difficulty	easy med. & hard	easy med. & hard	easy & med.
Amount	34 & 34 Qs	29 & 32 Qs	28 & 30 Qs
Limit	all	all	all
Strategy			
Content		no choice	
Question		after page	
Help		on error	
Ans-deducted		many tries at questions	

CBT courses

Two CBT courses were constructed from the courseware. They were 36 pages in length (from a potential 49) for CBT Genetics1 and 53 (from 74) were included in CBT Genetics2.

METHOD

Design

The study employed the same crossover design as Study One.

Participants

Sixty-six students from a local City Technology College were selected to take part in the study. These students were in three different classes grouped into ability sets. Unfortunately, a very substantial number of students were not able to complete the whole study. In particular, 40% of students were restreamed into classes not taking part in the study three weeks into the intervention. Secondly, the author who had originally been involved in creating the ITSs and who was class teacher to two of the groups gained a new job and left the school. The supply teachers who replaced him were unfamiliar with the material, experiment and students. These factors in addition to the standard problem of absences meant that there is full data from 15 students, 11 in the top ability set, and 4 in a lower ability set. They were between 14 and 15 years old and there were nine boys and six girls.

Materials

Pre and post-tests were identical to Study One except for changes needed to adapt them to the current curriculum.

Procedure

1. Pre-tests were given to the participants in their school classroom just prior to the intervention.
2. Intervention: The study was carried out at the school. Each session lasted either 45 or 90 minutes and there were a total of seven sessions. There was one experimenter and one teacher on hand to deliver non-computer tasks, provide help with the interface to the software and provide classroom management. Participants were provided with instruction booklets to help them navigate through the courses. No direct teaching of the concepts took place.
3. The post-tests were given to the participants within two weeks of their finishing the study.

RESULTS

Learning outcomes

To examine the effects of the intervention, a [2 by 2 by 2] ANOVA was carried out on the pre-test and post-test data. The design of the analysis was 2(Genetics1, Genetics2) by 2(pre-test, post-test) with a between-subjects factor of order of environment (REDEEMGenetics1/ CBTGenetics2, REDEEMGenetics2/CBTGenetics1) (Table 5).

Table 5
Pre and post test scores (out of 30) by course and type of environment

	REDEEM				CBT			
	Genetics1 (n = 7)		Genetics2 (n = 8)		Genetics1 (n = 8)		Genetics2 (n = 7)	
	\bar{x}	S.D.	\bar{x}	S.D.	\bar{x}	S.D.	\bar{x}	S.D.
Pre-test	12.57	5.16	12.88	5.57	11.38	5.01	14.00	3.21
Post-test	18.29	5.12	16.63	4.87	13.00	4.75	17.14	3.80

There was a significant main effect of time ($F_{1,13} = 54.39$, $MSE = 3.48$, $p < 0.001$) with post-test scores at 54% significantly higher than pre-test scores at 43%. The interaction between time, course and environment was significant ($F_{1,13} = 4.64$, $MSE = 4.43$, $p < 0.05$). Simple Main effects analysis showed that subjects' scores on Genetics1 and Genetics2 improved whether they learn the course with REDEEM or as CBT, except for those subjects who received Genetics1 as CBT (REDEEM Genetics1, $F_{1,13} = 32.82$, $MSE = 3.48$, $p < 0.0001$; REDEEM Genetics2 $F_{1,13} = 9.95$, $MSE = 3.48$, $p < 0.001$; CBT Genetics2 $F_{1,13} = 16.84$, $MSE = 3.48$, $p < 0.002$, but CBT Genetics1 $F_{1,13} = 3.04$, $p = ns$). Fig. 12 graphs this interaction as improvement in performance.

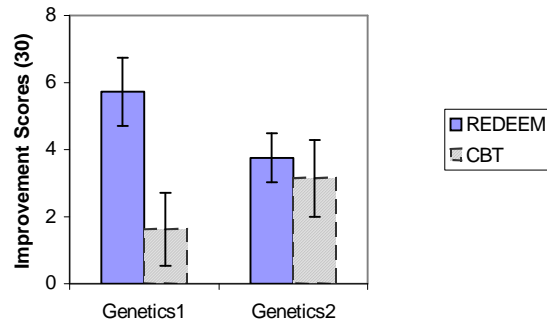


Fig. 12. Improvement scores by type of environment and course

To examine if students with higher prior knowledge learnt more, the relation between pre and post-test performance was examined. There was a significant positive correlation between pre-test scores and post-test scores ($r = 0.88$, $N = 15$, $p < 0.0005$), but no significant relationship between pre-test scores and improvement scores, ($r = -0.30$). This indicates that students at all levels of prior knowledge made similar improvements from pre to post-test. There was no relationship ($r = -0.17$) between students' improvement on the two courses.

Analysis of question type was performed using two [2 by 3 by 2] ANOVAs on the REDEEM and CBT data respectively, with two within-subjects factors, time and question type and one between-subjects factor, course.

Table 6
Pre and post test scores (out of 10) by question type, course and time (REDEEM)

Table 7
Pre and post test scores (out of 10) by question type, course and

	Genetics1(n = 7)						Genetics2 (n = 8)					
	Question type						Question type					
	RED/10		NT/10		Non/10		RED/10		NT/10		Non/10	
	\bar{x}	SD	\bar{x}	SD	\bar{x}	SD	\bar{x}	SD	\bar{x}	SD	\bar{x}	SD
Pre-test	4.29	2.29	4.71	1.89	3.57	1.72	4.25	2.05	4.63	1.51	4.00	2.56
Post-test	6.57	2.23	6.57	2.07	5.14	1.77	6.38	1.85	5.75	2.05	4.50	1.60

time (CBT)

	Genetics1 (n = 8)						Genetics2 (n = 7)					
	Question type						Question type					
	RED/10		NT/10		Non/10		RED/10		NT/10		Non/10	
	\bar{x}	SD	\bar{x}	SD	\bar{x}	SD	\bar{x}	SD	\bar{x}	SD	\bar{x}	SD
Pre-test	3.13	1.64	3.88	1.55	4.38	2.26	5.14	1.86	4.43	0.98	4.43	1.51
Post-test	3.75	1.83	4.50	2.39	4.75	1.67	6.57	1.51	5.43	1.99	5.14	1.35

For the REDEEM data there was a significant main effect of time ($F_{1,13} = 37.18$, $MSE = 1.50$, $p < 0.001$) and question type ($F_{2,26} = 6.70$, $MSE = 1.77$, $p < 0.005$) with both REDEEM and Near Transfer questions scoring significantly better than Non-REDEEM questions ($q = 4.4$, $p < 0.05$ and $q = 4.53$, $p < 0.01$ respectively) (Table 6). Interactions between question type and course, and between time and question type were not significant. The analysis of the CBT data showed a significant effect of time ($F_{1,13} = 12.44$, $MSE = 1.14$, $p < 0.005$) and the interaction between question type and course was also significant ($F_{1,13} = 5.37$, $MSE = 11.24$, $p < 0.02$) (Table 7). Simple main effects analysis showed that there was a significant difference between REDEEM questions as Genetics2 questions were answered significantly better at both pre and post-tests. There were no further differences between the conditions.

REDEEM and CBT process measures

The time on task data from school classroom is too noisy to be sensibly interpreted being affected by factors such as absence from classroom, messages from teachers, off task activities, *etc.* However, process data that recorded students' interaction with REDEEM system can still be examined (*e.g.* use of help, amount of notes).

Use of notes

Students failed to take up the opportunity to take notes when doing the CBT, so analysis was carried out on the REDEEM data only. On average, they wrote only 125.2 words (S.D. = 171.91) and there were no significant correlations between number of words written with REDEEM and any aspect of students' performance

REDEEM process measures

The author chose not to allow hints on request, so the only process measure analysed was students' performance on questions during intervention sessions. On average a total of 57% of answers were correct first time and there was a consistent relationship between question behaviour and students' prior knowledge. Those students who answered more question correctly first time scored higher on pre-tests ($r = 0.67$, $N = 15$,

$p < 0.01$) and post-tests ($r = 0.80$, $N = 15$, $p < 0.001$). This latter association remained even when pre-test scores were partialled out ($r = 0.58$, $N = 15$, $p < 0.05$).

DISCUSSION

Learning Outcomes

Students' performance at post-test was significantly higher than it was at pre-test. Moreover, this main effect was modified by the predicted three way interaction, *i.e.* that students scores would differentially improve for the course (either Genetics1 or Genetics2) they took with REDEEM. This is essentially what we observed. The analysis showed that Genetics1 scores were significantly higher when students learnt the material with REDEEM rather than as CBT. However, contrary to our prediction this was not true for Genetics2. This is plausibly an order effect as there are no obvious differences in authoring between the two courses. Students moving from REDEEM to CBT do slightly better than predicted and it is possible that the "good habits" are transferring. Those progressing from CBT to REDEEM do slightly worse than expected and may have less helpful interaction with REDEEM. For example, we found that students who took REDEEM Genetics2 wrote significantly less notes than those who did REDEEM Genetics1. A planned full cross over design (*i.e.* two further conditions of REDEEM/ REDEEM and CBT/ CBT) will provide more insight into these effects.

The impact on question type of performance is harder to analyse in this study than in Study One. REDEEM data did show the predicted effects of REDEEM and Near Transfer questions being answered better, but this effect was not modified by time. It would appear that by chance the students were less familiar with the Non-Redeem questions than with the other questions. It also appears in analysing the CBT data that Genetics2 questions (particularly REDEEM ones) were also more familiar to these participants. These factors make it difficult to account for any interaction between question type and environment on post-test performance.

Process measures

Students in the CBT condition did not use their pen and paper books to write notes. REDEEM students wrote only an average of 125 words per person. Perhaps because of this low value, there was no correlation between amount of notes and learning outcomes. There was a strong positive association between the percentage of questions answered correctly first time and performance. For example, there was a significant relationship with post-test scores, which remained even when pre-test scores were partialled out. Students who required fewer attempts to respond with the correct answer during the REDEEM sessions learnt more from the experience than those who required multiple attempts.

GENERAL DISCUSSION

Students in these studies improved their scores from pre-test to post-test. However, the main question of interest is whether learning with REDEEM led to greater improvement in these scores than learning with CBT. Across the two studies, the positive impact of REDEEM on learning outcomes varied greatly. For Study One, REDEEM scores only improved by 2% more than CBT scores. For Study Two, learning with REDEEM did lead to significantly greater improvement than learning with the CBT for Genetics1 as REDEEM scores improved by 20% whereas CBT improved by 5%. REDEEM in this case is substantially better than CBT (1.33 sigmas). However, this degree of difference was not maintained for Genetics2 with REDEEM scores being marginally better (0.31 sigmas). There was also an indication of potential benefit as questions on the post-test which had been included in the REDEEM's intervention showed the greatest degree of improvement (Study One) and that students who spent longer working with REDEEM learnt more (Study One). We were also happy REDEEM appeared to slow the students down, given the generally inappropriately low times spent with the learning environments.

Interpretation of these results requires overcoming a large credit assignment problem. Evaluating the effectiveness of an ITSAT requires consideration of many different interacting factors and as we argued above REDEEM is particularly problematic. Three factors that we have considered include 1) the authoring of the courses - in REDEEM's case, this primarily concerns the interactive features (*e.g.* informative questions, supportive feedback on answers, helpful hints, reflection points at appropriate places) and course structure. 2) The macro-adaptation features – was differentiation based upon an appropriate classification and were appropriate teaching decisions made based upon this classification? 3) An ITS delivered with REDEEM also depends on external underlying courseware.

1) Both teachers had detailed knowledge of the topic, which allowed them to create a clear domain structure and to provide questions and exercises on issues that were judged to be both important and likely to be difficult. Some independent evidence of this is seen from the minimal changes that the second teacher made to questions, reflection points and non-computer-based tasks that the first author created. Some of the hints were rewritten and the course was reorganized when he created learning environments for the second study. But, it would be difficult to argue that there was anything qualitatively better about the authoring of the second Genetics1 learning environment.

2) The basis of the teachers' classification of their students into different categories also seemed fairly unproblematic. In Study One, the teacher had a detailed knowledge of the students, which was evident in the way that her judgments of their knowledge of Genetics correlated highly with their pre-test scores. In the second study, the teacher did not determine categorisation but implemented a previously agreed streaming procedure that was monitored by the school. These categories were used differently by the teachers with respect to REDEEM's differentiation features. In Study One, student categories were assigned really quite different material (see Ainsworth, *et al*, 2002) but there was little difference between the content for different categories for Study Two. The teacher expressed dissatisfaction with this, saying he was including material he felt some students had little chance of understanding but he was forced to use it because of the syllabus. In both studies, the researchers would have tended to make some teaching strategy decisions differently to the authors. For example, in Study One the author chose to limit help on request and fixed fairly strict number of attempts at questions for some groups. In Study Two, we would have assigned different strategies to these different groups. For example, in line with the research on the relation between prior knowledge and learner control (*e.g.* Chung & Reigeluth, 1992) we would have allowed Group 1 students more control over their learning. However, this is just opinion as there is no independent evidence as to whose views on teaching strategy were the most appropriate for these particular students. Furthermore, it is unlikely that this could explain the majority of the results. In Study One, all groups of students improved similarly from pre to post-test and each group received a different teaching strategy. A more fundamental question that we will come back to later is whether the alternative teaching strategies that REDEEM allows could impact significantly on different students' learning..

3) Unlike other ITS authoring tools, REDEEM relies on the pre-existing content. If this underlying courseware is not of good quality, then REDEEM may be unable to do much to enhance it. Alternatively, if the courseware is already rich in interactivity and allows for students with different needs then perhaps REDEEM's feature are superfluous. The courseware explanation does not appear likely. The CBT had already been used in a school classroom and contains much clearly presented and relevant information. However, given its limited interactivity and scope for different interpretations of the material, it could still be enhanced by REDEEM's features. Again, there is no reason why the second Genetics1 would have been differentially enhanced by REDEEM.

However, there are two other factors that are worth considering when investigating why REDEEMed environments appeared to be beneficial in some circumstances and not in others, namely the wider context of the studies and the learners.

Both studies involved students who had a genuine need to learn the material as part of their education and the material they studied had been deemed suitable and relevant by their class teachers. As the authors of the learning environments were their class teachers, they had detailed knowledge of both the domain and the students. However, in Study One, students came to the University to study, which reduced the authenticity of the study. To try to address this issue in Study Two REDEEM was used in the school classroom. Unfortunately, student drop out seriously compromised this study. Furthermore, practical problems made it

difficult to teach the material in the time available. Setting up laptops at the beginning of each lesson was time-consuming and was done in a different room for each class with little or no breaks between lessons. We therefore lost around 25% of each lesson to this activity. These issues are worth rehearsing as they represent the continuum of problems facing evaluations of learning environments. Experiments under laboratory conditions allow researchers reasonable control over variables and process data but are artificial and outside learners' everyday experience. Whereas authentic evaluations in actual classrooms provide much less control over the setting and lead to subsequent difficulties of interpreting noisy data. We don't believe there is a way solve these problems and so remain committed to trying to do both wherever possible.

The final aspect of the studies that is worth considering is the role of the students in these studies. One reason why these differences between the results for CBT and REDEEM may be lessened is that individual students may adjust to differing environments to maintain their performance. Both environments deliver the same (apart from differentiated content) declarative material. Consequently, whilst it may be easier to learn this material when you are interacting with a system that asks you questions, provides hints to their solution, provides you with an on-line note tool, *etc*, it is of course still possible to learn without these facilities. Thus, students could compensate for the lack of support in the CBT by working harder. However, this explanation is not supported by the data. There was no correlation between students' improvement scores on their REDEEM course and their CBT course. In contrast, we need to acknowledge significant problems with participants' motivation. Many, though by no means all of the students, did not wish to learn about this topic. This was evident from the general time spent on reading material and interacting with exercises. REDEEM provides student history inspection tools and it was obvious that many students were skipping through pages without reading them. The trace logs from the CBT if anything revealed an even worse picture. This was also more noticeable with Genetics2 than Genetics1. This was true in both studies but affected the results of the experiment differently. In Study One, all students completed the intervention and their data were included in the analysis, whereas in Study Two motivation to participate was so low that in Groups 2 and 3 the vast majority of students did not finish and so were excluded from the analysis. Hence, Study Two that found that REDEEM's increased learning relative to the CBT excluded the students that were particularly low in motivation. REDEEM may provide more features that support learning, but learners need to engage with the system if they are to benefit from those features. Hints are only helpful if you read them, exercises only beneficial if you complete them and on-line note tools only valuable if you write in them. The significant correlations between amount of notes written, percentage of questions answered correctly first time and time spent learning with REDEEM show that unsurprisingly students who took advantage of REDEEM's features learnt more than those who did not.

Exploring the difference between the REDEEM learning environments and the CBT

REDEEM differs from the underlying courseware in three ways: 1) the structure of material, 2) the interactive features, and 3) the macro-adaptation features. In these studies, the classroom teacher asked the researchers to create a structure for the CBT courses, which while they differed to the REDEEM learning environments, were adapted to their view of teaching. As a consequence, we reduced the difference between traditional use of CBT and REDEEM. Therefore the main differences between the two environments lie in the interactive and the macro-adaptation features. We had hypothesised that increasing the interactivity of the environment would lead to better learning for all students. We also proposed that by adapting the teaching styles and content to specific learner groups, we would also improve learning.

The results of the studies suggest that any observed advantage of REDEEM was due more to interactivity than macro-adaptation. The evidence for this statement comes firstly from the fact that in Study One, which differentiated content and strategies far more than Study Two, we observed no difference in learning outcomes between REDEEM and CBT or between any of the REDEEM learning environments. Secondly, the process measures that identified the students who had made the greatest improvement showed that it was those students who interacted most with the environment. That degree of interaction predicted learning outcomes does not seem contentious but the question that remains is why we did not observe benefits from macro-adaptation. We propose three potential reasons: 1) macro-adaptation is unimportant; 2)

that the macro-adaptive strategies used in this study were inappropriate; and 3) that macro-adaptation was potentially appropriate and important but that other factors inhibited its impact.

1) The research on aptitude-treatment interactions and ITS design suggests that macro-adapting the teaching strategy used by the environment should lead to better learning than using one strategy for all learners. For example, Arryo *et al* (2000) showed that macro-adapting hint style by gender and level of cognitive development was beneficial and Shute (1992) showed that the explicitness of feedback should be adapted to learner's ability. Although much research on ATI has shown null to slight results (*e.g.* Cronbach & Snow, 1977), given the level of control possible with an ITS, it seems plausible that we should expect macro-adaptation with ITSs to deliver more clear-cut results than in classroom situations. Therefore, it seems unlikely that macro-adaptation *per se* is not effective.

2) A further explanation is that authors' made inappropriate decisions about categorisation or about how content or strategy was adapted to these categories. However, there is little doubt that the teachers' categorisation of students, which was based on their perceived aptitude, was accurate. However, aptitude may not be the most appropriate dimension to use to rank the students or it might need to have been supplemented by other factors. For example, some students in Group A and B, who had high control over when they answered questions, did not use the feature appropriately. Some chose not to answer the questions until they found that REDEEM would not let them quit without doing so. When talking to the teacher about this, she was unsurprised, commenting that she felt that certain students in this category were not as highly motivated as she would like. Potentially, factors such as motivation and self regulatory skills could be used to set teaching dimensions such as level of student control of material, questions and help seeking and aptitude used to set difficulty of material and questions. REDEEM can easily accommodate this approach but to date no teacher has chosen to use any factor other than familiarity with the content and aptitude as classifying variables.

So assuming the classification was appropriate and accurate, did teachers make the best decisions about how to assign content and strategies? From inspection of the learning environments, there is little doubt that the Groups A and B saw more complex material and answered more difficult questions than Groups C through E. However, it is not possible to independently determine if this was appropriate to their needs. The percentage of question answered correctly first time was significantly higher in Group A than D and E which could be viewed as indicative that one or more groups were getting questions that were inappropriately easy or difficult. If questions had truly been adjusted to aptitude, there should have been equal performance across all categories. However, we believe that answering questions correctly first time indicates not just prior knowledge and learning, but also the level of students' attention and effort.

We should also examine the teacher's decisions about macro-adapting REDEEM's teaching strategy to student category. On the whole, her views on teaching are generally in line with the ATI literature. For example, she tended to use more learner control in higher groups, in line with research that finds that those students who score higher on pre-tests learn better with more autonomy (*e.g.* Williams, 1996). She also provided higher ability students with more opportunities to induce their own answers to the questions. This is consistent with Shute (1992) who found that higher ability subjects learned more declarative knowledge in rule-induced environments and lower ability subjects learned more in rule-given environments. However, there may be more disagreement between the author's decision and previous findings with help seeking, as higher-scoring students were not allowed help on request. There is some evidence that higher-scoring learners may be better able to judge when they should seek help (*e.g.* Wood & Wood, 1999) and certainly no evidence that they request help unnecessarily or are help-abusers. But, overall there is little indication that the macro-adaptation was inappropriate and quite a bit of evidence that it was appropriate.

3) The teacher's use of REDEEM's content and teaching strategy adaptation features are primarily in line with that of the research literature. However many other variables than prior performance have been explored (*e.g.* learning style, working memory capacity, self-regulatory skills, visualiser/verbaliser, general knowledge, gender, high anxiety/low anxiety, level of cognitive development) and potentially these should have been included. Furthermore, many of these factors may show up in laboratory studies but their effect size may be somewhat low and they may have little impact in the classroom. The number of students in each category in Study One would often not have been sufficient to allow identification of benefits unless they were very substantial. Moreover, such benefits might be difficult to identify. For example, if an author

assigned a unique teaching strategy to every category of learner and they all made equal gains, does this mean that the strategies were ideally targeted or that they had no effect? Consequently, we need further research to examine the educational significance of macro-adaptation and to consider which are the most important learner characteristics and strategy dimensions. It remains an open question as to the cost and benefits of performing pre-tests and evaluation of learners in relation to the impact of macro-adaptation. Shute (1993) argues it may be relatively cost-effective to implement testing of such factors as working memory and then change environments as a result. However, much greater problems lie determining what factors should be pre-tested and what features of environments should change, let alone the problem of combinations of characteristics (*e.g.* a high WM male, with poor self-regulatory skills, a great deal of familiarity with the material, low anxiety and a preference for visual material).

CONCLUSIONS

In two experiments, the relative effectiveness of REDEEM learning environments and CBT were compared. Analysis revealed that there was an advantage for REDEEM ITSs in terms of learning outcomes but the effect size was highly variable ranging from 0.1 to 1.33 (mean 0.51). By examining process measures and authoring decisions, we argued that REDEEM's primary benefit in these studies was the way it easily allowed teachers to add extra interactivity. It was those students who took advantage of the interactive features (*e.g.* by answering question and writing notes) who gained the most from the experience.

If REDEEM had reliably generated the degree of improvement we saw for Genetics1 in study two, then there would be little argument about whether the time needed to author with REDEEM was cost-effective. In addition, REDEEM does offer significant advantages for classroom use. Firstly, we showed that REDEEM improved learning for those students who were prepared to engage with its interactive features. REDEEM keeps detailed student histories that are used as the basis of learner, class or course reports. It is therefore easy to see from the reports who is not interacting with the system. This could also be automated. Secondly, although we have found little evidence in these studies that macro-adapting REDEEM to different student categories led to significantly better learning outcomes, teachers welcomed the opportunity to quickly adapt a course to learners' needs. This may increase the chance of such software being practical in the complex classroom environment. Thirdly, these studies used macro-adaptation with respect to ability categories. However, the capability to change teaching strategy means that REDEEM can take the same material and adapt it to different functions as easily as different learners. One real possibility is to use strategies that are developed for functions such as whole class presentation, initial exploration by learner, revision, *etc.* It is also possible to explore other learner characteristics (such as motivation or self-regulatory skills) as the basis of macro-adaptation.

Currently, we are exploring the basic ideas behind REDEEM in a number of different contexts. We are considering whether learning outcomes could be improved by increasing the intelligence of ITSs for example, by including more micro-adaptation functions. In particular, we would like to increase the sophistication of questioning and associated remediation as ablation studies (*e.g.* Shute, 1995) identify the importance of these features. This is particularly important if REDEEM is used in problem-solving domains. To date, most of material taught with REDEEM has been declarative in nature.

However, this raises an interesting dilemma as teachers have a tendency to try and use REDEEM in such a way as to make the learning environments less smart (*e.g.* by attempting to prescribe a strict prerequisite structure, using fixed not performance related categories). Should we encourage teachers to envisage a future where they work in collaboration with more intelligent software or instead provide them with tools to create software to fulfil their requirements today? We are also exploring the role of the learner in authoring environments. For example in a version of REDEEM for University courses, we have implemented a mixed initiative model where authors make decisions about course structure/content and interactivity but students choose how to macro-adapt REDEEM to their own personal preferences. Thus, in keeping with Murray's (2003) call for meta-authoring tools we envisage a future of multiple REDEEM authoring tools that vary in the complexity of the authoring and the flexibility of resulting systems. Such an

approach acknowledges that just as different learning environments must be adapted to different contexts so must the authoring tools that create them.

ACKNOWLEDGEMENTS

This research was supported by the ESRC at the ESRC Centre for Research in Development, Instruction and Training. We are very grateful to the teachers and schools without whose help this research would not have been possible. We would also like to recognize the assistance of other members of the Centre in running these experiments, particularly Jo Cheng, Nigel Pitt, Ben Williams and Heather Wood. Over the years, a number of people have contributed to the REDEEM project, especially Jean Underwood and David Wood. Rose Luckin, Piers Fleming and Jean Underwood provided helpful comments on the first draft of this paper. Finally, we would like to acknowledge Nigel Major, without whom none of this would ever have happened.

REFERENCES

- Ainsworth, S., Bibby, P., & Wood, D. (2002). Examining the effects of different multiple representational systems in learning primary mathematics. *Journal of the Learning Sciences*, 11, 25-61.
- Ainsworth, S., Grimshaw, S., & Underwood, J. (1999). Teachers implementing pedagogy through REDEEM. *Computers & Education*, 33(2-3), 171-187.
- Ainsworth, S., Underwood, J., & Grimshaw, S. (1999). Formatively evaluating REDEEM - An authoring environment for ITSs. In S. Lajoie & M. Vivet (Eds.) *Proceedings of AI-ED 99*, (pp. 93-100). Amsterdam: IOS.
- Ainsworth, S., Underwood, J., & Grimshaw, S. (2000). Using an ITS authoring tool to explore educators' use of instructional strategies. In G. Gauthier & C. Frasson & K. VanLehn (Eds.) *Proceedings of Intelligent Tutoring Systems 2000* (pp. 182-191). Berlin: Springer-Verlag.
- Ainsworth, S., Clarke, D., & Gaizauskas, R. J. (2002). Using edit distance algorithms to compare alternative approaches to ITS authoring. In S. A. Cerri & G. Gouardères & F. Paraguaçu (Eds.) *Proceedings of Intelligent Tutoring Systems 2002* (pp. 873-882). Berlin: Springer-Verlag.
- Ainsworth, S., Williams, B., & Wood, D. (2001). Using the REDEEM ITS authoring environment in naval training. *Proceedings of the IEEE International Conference on Advanced Learning Technologies* (pp. 189-192).
- Ainsworth, S., Major, N., Grimshaw, S. K., Hayes, M., Underwood, J. D., Williams, B., & Wood, D. J. (2003). REDEEM: Simple Intelligent Tutoring Systems From Usable Tools. In T. Murray & S. Blessing & S. Ainsworth (Eds.) *Tools for Advanced Technology Learning Environments*. (pp. 205-232). Amsterdam: Kluwer Academic Publishers.
- Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26(2), 147-179.
- Arroyo, I., Beck, J. E., Woolf, B. P., Beal, C. R., & Schultz, K. (2000). Macro-adapting Animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. *Proceedings of the 5th International Conference ITS 2000* (pp. 574-583). Berlin: Springer-Verlag.
- Bell, B. (1998). Supporting educational software design with knowledge-rich tools. *International Journal of Artificial Intelligence in Education*, 10, 46-74.
- Blessing, S. B. (1997). A programming by demonstration authoring tools for model tracing tutors. *International Journal of Artificial Intelligence in Education*, 8(3-4), 233-261.
- Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A metaanalysis of findings. *American Educational Research Journal*, 19, 237-248.
- Cohen, P. R., & Howe, A. E. (1988). How evaluation guides AI research. *AI Magazine*, 9, 35-43.
- Corbett, A. T. & Anderson, J. R. (1991). Feedback control and learning to program with the CMU LISP tutor. Paper presented at the AERA, Chicago, IL.
- Chung, J. & Reigeluth, C.M. (1992). Instructional prescriptions for learner control. *Educational Technology* 32(10), 14-20
- Cronbach, L.J. & Snow, R.E. (1977). *Aptitudes and Instructional Methods*. New York: Irvington.
- du Boulay, B. (2000). Can we learn from ITSs? In G. Gauthier & C. Frasson & K. VanLehn (Eds.) *Proceedings of Intelligent Tutoring Systems 2000*, (pp. 9-17). Berlin: Springer-Verlag.
- Graesser, A. C., Person, N. K., Harter, D., & The Tutoring Research Group. (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, 12, 257-279.

- Hsieh, P. Y., Half, H. M., & Redfield, C. L. (1999). Four easy pieces: Development systems for knowledge-based generative instruction. *International Journal of Artificial Intelligence in Education*, 10, 1-45.
- Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8, 30-43.
- Lesgold, A., Lajoie, S., Bunzo, M., & Eggan, G. (1992). Sherlock: A coached practice environment for an electronics troubleshooting job. In J. Larkin & R. Chabay (Eds.) *Computer Based Learning and Intelligent Tutoring* (pp. 202-274). Hillsdale, NJ: LEA.
- Luckin, R., & du Boulay, B. (1999). Ecolab: The development and evaluation of a vygotskian design framework. *International Journal of Artificial Intelligence in Education*, 10, 198-220.
- Major, N.P. (1994) Evaluating COCA - What do teachers think? *Proceedings of the World Conference on Educational Multimedia and Hypermedia - EDMEDIA 94*. AACE Press.
- Major, N., Ainsworth, S. E., & Wood, D. J. (1997). REDEEM: Exploiting symbiosis between psychology and authoring environments. *International Journal of Artificial Intelligence in Education*, 8(3/4), 317-34.
- Mark, M., & Greer, J. E. (1995). The VCR tutor: Effective instruction for device operation. *The Journal of the Learning Sciences*, 4(2), 209-246.
- Meyer, T. N., Miller, T. M., Steuck, K., & Kretschmer, M. (1999). A multi-year large-scale field study of a learner controlled intelligent tutoring system. In S. Lajoie & M. Vivet (Eds.), *Proceedings of AI-ED 99* (pp. 191-198). Amsterdam: IOS.
- Munro, A., Johnson, M. C., Pizzini, Q. A., Surmon, D. S., Towne, D. M., & Wogulis, J. L. (1997). Authoring simulation-centered tutors with RIDES. *International Journal of Artificial Intelligence in Education*, 8(3-4), 284-316.
- Murray, T. (1997). Expanding the knowledge acquisition bottleneck for intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 8(3-4), 222-232.
- Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education*, 10, 98-129.
- Murray, T. (2003). An overview of intelligent tutoring system authoring tools: Updated analysis of the state of the art. In T. Murray & S. Blessing & S. E. Ainsworth (Eds.) *Tools for Advanced Technology Learning Environments*. (pp. 491-544). Amsterdam: Kluwer Academic Publishers.
- Murray, T., & Woolf, B. (1992). Tools for teacher participation in ITS design. In C. Frasson & G. Gauthier & G. I. McCalla (Eds.) *Proceedings of Intelligent Tutoring Systems 92* (pp. 593-600). Berlin: Springer-Verlag.
- Russell, D.M., Moran, T.P. & Jordan, D.S. (1988) The instructional design environment. In J. Psotka, S.A. Massey & A. Mutter (Eds.) *Intelligent tutoring systems: Lessons learned*. Hillsdale, NJ: LEA.
- Shute, V. J. (1992). Aptitude-treatment interactions and cognitive skill diagnosis. In J. W. Regian & V. J. Shute (Eds.) *Cognitive Approaches to Automated Instruction*. Hillsdale, NJ: LEA.
- Shute, V. J. (1993). A macroadaptive approach to tutoring. *Journal of Artificial Intelligence in Education*, 4(1), 61-94.
- Shute, V. J. (1995). SMART evaluation: Cognitive diagnosis, mastery learning and remediation. In J. Greer (Ed.) *Proceedings of AI-ED 95* (pp. 123-130). Charlottesville, VA: AACE.
- Shute, V. J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments*, 1, 51-77.
- Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, present and future. In D. Jonassen (Ed.) *Handbook of Research on Education Communications and Technology* (pp. 1 - 99). New York: Macmillan.
- Towne, D. M. (1997). Approximate reasoning techniques for intelligent diagnostic instruction. *International Journal of Artificial Intelligence in Education*, 8(3-4), 262-283.
- Williams, M. D. (1996). Learner-control and instructional technologies. In D. H. Jonassen (Ed.) *Handbook of Research on Education Communications and Technology* (pp. 957-982). New York: Simon & Schuster . Simon & Schuster.
- Woolf, B.P. & Cunningham, P.A. (1987) Multiple knowledge sources in intelligent teaching systems. *IEEE Expert*, Summer 1987.
- Wood, D., Bruner, J., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17, 89-100.
- Wood, D. J., & Wood, H. A. (1999). Help seeking, learning and contingent tutoring. *Computers & Education*, 33(2-3), 153-1770.