

**EVALUATING THE EFFECTIVENESS AND EFFICIENCY  
OF THE REDEEM INTELLIGENT TUTORING SYSTEM AUTHORING TOOL**

**Shaaron Ainsworth and Shirley Grimshaw**

ESRC Centre for Research in Development, Instruction & Training

School of Psychology, University of Nottingham

University Park, Nottingham NG7 2RD, United Kingdom

Email: sea@psychology.nottingham.ac.uk

October 2002

**ABSTRACT**

The REDEEM authoring environment allows teachers to create simple ITSs from existing CBT by imposing their pedagogical preferences about how different groups of students should best be taught. Consequently, a fundamental assumption behind the research is that students will learn either more effectively or more efficiently from these ITSs than from the original CBT. We conducted two studies with 14-16 yr old students learning genetics to test this assertion. In both experiments, a class teacher who had expert knowledge of both the topic and of the students constructed two ITSs with REDEEM from pre-existing CBT. Using a crossover design, the learning outcomes for students who studied these two courses (either a REDEEM then CBT course or *vice versa*) were compared. In the first study, we found that performance of the students (N = 86) improved from pre-test to post-test but learning outcomes were not influenced by type of learning environment. However, inspection of the process data revealed that students who engaged with REDEEM's features did learn more. In a second study, conducted in a more naturalistic context, a further 15 students completed the courses. The results of this study revealed that REDEEM could significantly improve learning compared to CBT. Detailed analysis of participants' performance suggested that REDEEM could enhance knowledge by supporting additional interactivity but that features such as macro-adaptation did not appear to impact upon performance. Possible interpretations of these results are discussed in the light of the many evaluation issues for ITS authoring tools.

**Keywords:** ITS authoring tools, evaluation, macro-adaptation

**INTRODUCTION**

Research has shown that when learners interact with Intelligent Tutoring System (ITSs), they can stimulate impressive learning outcomes (*e.g.* Koedinger, Anderson, Hadley & Mark, 1997; Mark & Greer, 1995; Lesgold, Lajoie, Bunzo, & Eggan, 1992). Yet despite these benefits, ITSs are only just beginning to achieve application in schools, colleges or workplaces. One potential reason for this has been the difficulty in developing such systems even in a relatively limited domain - creating an ITS is estimated to take 200-1000 hours to produce an hour of instructional material (*e.g.* Woolf & Cunningham, 1987; Murray, 1999). Consequently, one of the primary goals motivating the

development of ITS Authoring Tools (ITSATs) is to deliver the benefits of ITSs 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, 1995) 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 representations that authoring tools required them to manipulate. 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 ITS development. REDEEM represents one end of the continuum of the current generation of ITSATs (for a detailed review of other approaches, see Murray, 1999). REDEEM does not support the construction of domain material, but focuses on the authoring of pedagogy. Authors (who need not have any programming knowledge) import pre-existing domain material and then authoring tools capture their knowledge of how they want to teach this material. This allows the REDEEM shell to teach students in a way that is adapted to a learner's individual needs. REDEEM generated ITSs are among the least sophisticated of such systems. They have little domain knowledge compared to systems such as Demonstr8 (Blessing, 1997), which has a detailed production system account of the domain or Diag with its knowledge of fault finding and diagnosis (Towne, 1997). They rarely include complex simulations of the sort that RIDES supports (Munro, Johnson, Pizzini, Surmon, Towne, & Wogulis, 1997). The REDEEM tools themselves are generic in terms of the domains to which they can be applied but to achieve this we have sacrificed the benefits of knowledge rich tools (Bell, 1998). It can be seen that REDEEM is designed to offer teachers significantly educational flexibility without requiring complex and time-consuming authoring. Essentially, the comparison is between offering teachers the tools to write a whole textbook (traditional ITS authoring) versus giving them tools to customise a textbook to their particular needs (REDEEM). One primary reason for the experiments reported in this paper was to examine if this trade-off has been productive, i.e. can REDEEM allow teachers to create ITSs that are effective for their students.

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. So far, probably the only ITSAT that can claim large scale empirical evaluation is XAIDA (Hsieh, Half, & Redfield, 1998). One reason for the small number of evaluations of ITSATs relative to ITSs is that determining the success of an ITSAT is more complicated given the need to evaluate the authors' as well as the learners' experiences (e.g. Murray, 1997; Ainsworth, Grimshaw & Underwood, 1999).

If the evaluation goal is not simply to determine whether an ITS can promote successful learning outcomes but also to identify what features lead to this success, then establishing a causal relationship between aspects of the ITS design and positive learning gains is very complicated. To be done successfully, precise and large-scale experiments are required (*e.g.* Shute, 1992, Mark & Greer, 1995). Furthermore, the effectiveness of any learning environment is influenced by its context of use (see Wood, Underwood & Avis, 1999) and so evaluations need to be conducted in real situations (*e.g.* Koedinger, Anderson, Hadley & Mark, 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 its ITSs not only depend on the user, authoring tools and ITS shell, they also involve externally imported domain material. Consequently, no single study or approach is able to answer all questions about the benefits of an ITSAT.

In this paper, we present two studies that investigate whether REDEEM, in the hands of expert teachers, leads to more efficient or to effective learning. Ideally, learners interacting with REDEEM will come to understand the subject matter more completely, will have found the experience of learning motivating and should reach the desired learning goals in a more time efficient way. To try and determine the effectiveness of an ITS, experimenters have used a number of different alternatives as control (for a review see, du Boulay, 1999). 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 (2 sigmas) 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 ITSs would fall around the 0.5 range.

Another common technique is to compare learning environments to alternative versions of themselves (*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 or whether there are Aptitude-Treatment interactions (ATI) where one type of learner may benefit 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 in how many problems the tutor required students to complete. She also assessed students' general knowledge and working memory (WM) capacity. Overall, there were no differences in learning outcomes associated with the type of tutor, but there were ATIs such that High Knowledge, Low WM students learnt better with few problems and Low Knowledge, High WM students learnt better with many

problems. This approach aims to explore if additional benefits can be found from macro-adapting the style of a learning environment to an individual learners needs.

A variation on this technique is ablation experiments where particular design features are removed and performance of the 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 version of a tutor that taught learners how to operate a video recorder. The “smartest” one used model tracing to monitor learners’ performance and could therefore tutor for misconceptions with feedback that could be procedural and conceptual. The “dumb” version allowed the user only a one way to perform a task and provided only simple prompting. Learning with smarter systems decreased the number of steps, errors and time required for students to complete the post-test. Shute (1995) used this technique to assess the contribution the remediation component of Stat Lady, which detect is learners are not succeeding at a curriculum element and adjust instruction accordingly. She found that by enabling it 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 type of help offered and abstraction of terms. One version of the system was “dumb” having a limited student model and relying on learner’s insight into his or her own needs. 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 of the environment.

Previous studies with REDEEM have focused on the experience of authors in domains as diverse as “Shapes” in primary mathematics and Communication and Information System Principles” with the Royal Navy (see Ainsworth, Grimshaw & Underwood, 1999; Ainsworth, Williams & Wood, 2001). On average, only 90 minute are required to train authors. The majority of the tools are simple to use, especially those used for macro-adaptation. Authoring with REDEEM is time efficient. Authors have taken between 1 and 5 hours to develop an hour of instruction. So we have demonstrated that ITS production with REDEEM is relatively efficient but the key question of whether it is effective had not been addressed.

To answer this question, we recruited secondary (high) school teachers and provided them with 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 pupils who subsequently 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 ITSs. If learning outcomes are higher with REDEEM, then we can conclude that the REDEEM/Author partnership in the situation provided better support for learning than the non-intelligent courseware. If

authors create different ITSs for learner groups, then we may also be able to examine what aspects of teaching strategies are beneficial for (particular groups of) learners. Thus, in these studies we performed both an ablation experiment (stand-alone courseware versus REDEEM ITSs) and a within-system comparison (Multiple REDEEMs with differing teaching strategies). Before describing these experiments, we present a brief overview of how REDEEM creates simple ITSs.

## 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 (Figure 1). Authors use the REDEEM tools to describe courses, supplement them with learning activities, construct teaching strategies and identify particular 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 in turn. For a fuller description, the reader is referred to Major, Ainsworth & Wood (1997).

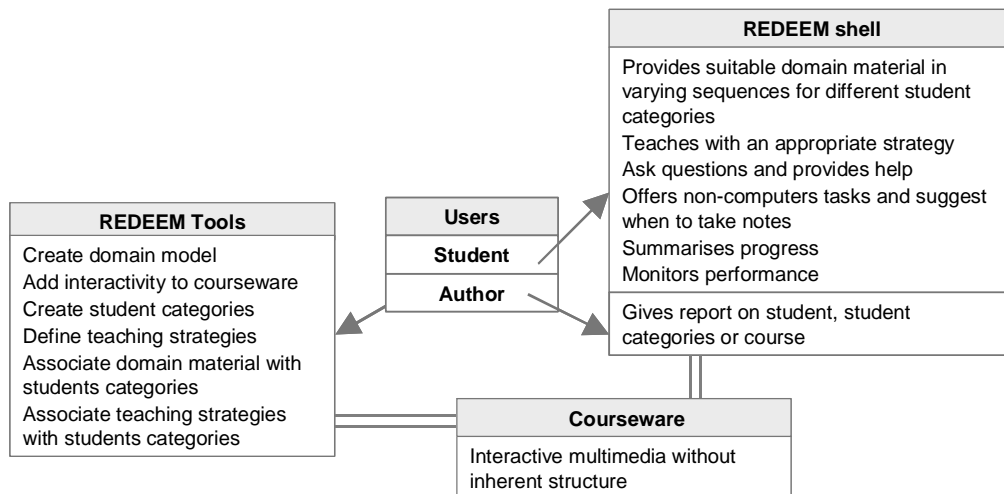


Figure 1. REDEEM schematic

## Courseware Catalogues

Domain material in REDEEM is based on the idea of a courseware catalogue. It consists of pages from computer-based training (CBT) developed in a standard authoring package, Click2Learn ToolBook or downloaded from the Internet. These are not built within REDEEM but are used to provide the basic pre-prepared subject content. Consequently, this 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. Pages can contain multi-media displays, simulations, animations, questions and exercises. REDEEM does not model the learners' actions on these pedagogical objects.

### 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 learner groups.

### What to Teach

One of the most important stages in the authoring process 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 find compatible with their teaching preferences.

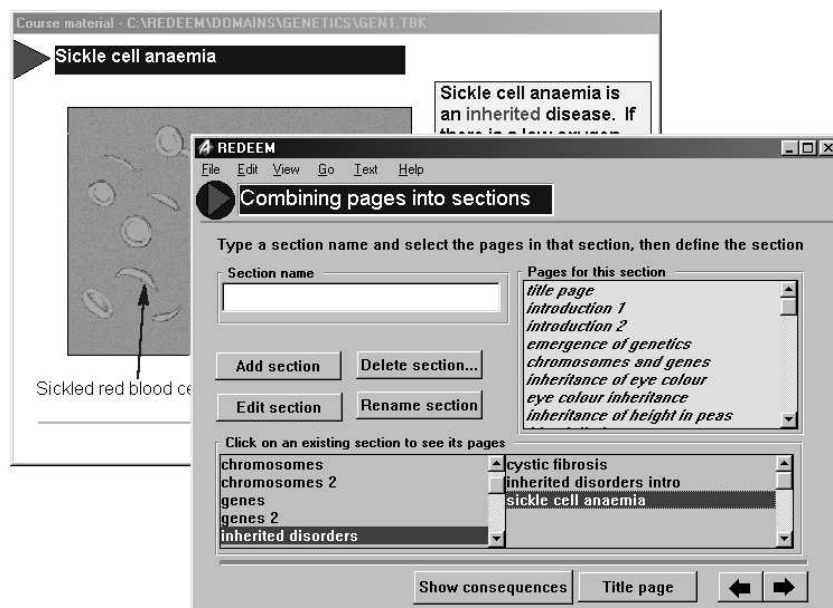


Figure 2. Creating sections and describing pages

Pages are combined into sections (see Figure 2). Pages can be placed in multiple sections and sections need not consist of contiguous pages in the underlying CBT. Sections are then described upon a number of dimensional ratings by using sliders. For example, authors describe how familiar or easy a section is likely to be for their students. In addition, authors may describe relations between sections, such as “pre-requisite” or “applies”. Pages themselves are then described in terms of the same dimensional ratings and relations, but 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. There are only 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. Firstly, authors can associate a reflection point with a page, so the ITS shell will suggest that students should consider taking notes by displaying an on-line notes tool. Secondly, authors can associate a non-computer task with the page. This ensures that students are directed to leave REDEEM and perform an appropriate learning activity. Thirdly, authors can create questions (multiple choice, fill in the blank, multiple true, true-false and matching questions) and provide feedback that will explain to the student why an answer is correct. 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, which ideally increase in specificity. Finally authors describe a number of characteristics of the question that the ITS shell uses to decide how to use the question given a specific teaching strategy. Thus, 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 whether its position is held constant or whether it can vary with the teaching strategy.

### **Who to Teach**

Students are described as belonging to one of a set of teacher-defined categories. These can be at any degree of granularity ranging from the whole class to an individual child. 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. If this is the case, then 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. Figure 3 shows the teacher defining a "practice" strategy. A graphical interface allows teachers to choose a number of potential teaching and testing styles.

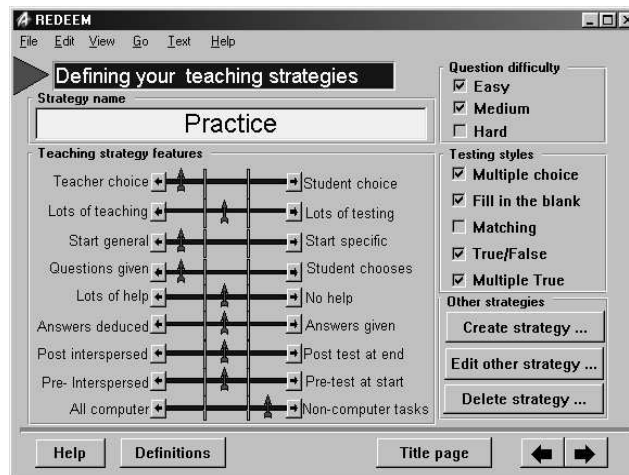


Figure 3. Creating a teaching strategy

Different instructional principles can be embodied in various strategies by manipulating the sliders. Each slider has three discrete positions that result in different instruction. Consequently teachers are free to create as many strategies as want 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 (*e.g.* exclude matching questions). Teachers can create as many strategies as they wish. In fact, REDEEM can offer over 10000 different teaching strategies each subtly different to each other, although in the studies conducted so far, no author has created more than seven.

### What Students Learn

Authors differentiate material for groups of learners by associating sections of the course with student groups. By default, learners see all the material, but the teacher can choose to remove sections for a particular category. This is commonly used by authors to focus on introductory material for learners who need more help whereas higher performers group may be given more difficult sections to allow them to spend more time on more complex aspects of the course. This can be achieved by the use of alternative sections that cover roughly the same material but at different levels of difficulty.

### How Students Learn

The final necessary stage of authoring is to relate the each student category to a teaching strategy. To date, authors have varied from creating a single preferred teaching strategy to creating a unique strategy for each group. Authors have tended to focus either on the perceived knowledge of their students, their perceived abilities or their role (first time versus reviser).

### ITS Shell

The ITS shell delivers the courseware according to the instructions generated by teachers using the authoring tools in combination with its predetermined defaults. To do so it includes a limited student model, which also serves the basis of reports given to teachers.

### Delivering adaptive instruction

The main role for the ITS shell is to deliver the course material to each student in the manner that the teacher has specified using the authoring tools (Figure 4). 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 on 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.

If the shell is teaching, it computes a weighted array of choices using the semantic network of pages and its default assumptions. Such assumptions include prefer easy material before difficult material, introductory before final 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 pre-requisite 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 the 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 differs depending on the strategy. Hence, it computes the number of attempts per questions a learner is allowed and whether to offer help either on request or error.

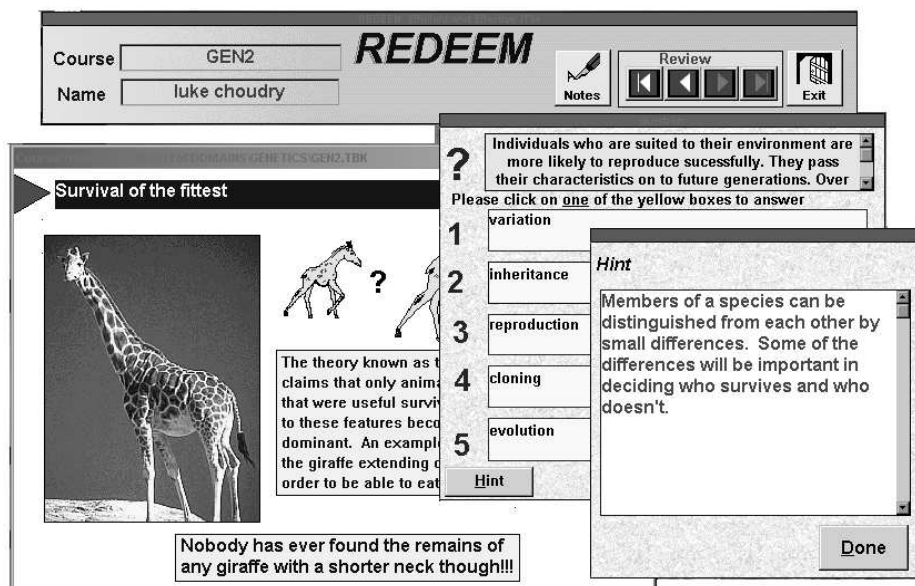


Figure 4. REDEEM Shell running Genetics2

### **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. The basic course material unit being modelled is the page. The shell maintains a student history in addition to the student model. 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 question right, using help appropriately, *etc.* Student category reports allow teachers to compare performance of a group of students, for example, 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 to check on the way that particular questions are answered to assess if they are at the appropriate level of difficulty.

## **STUDY 1**

### **Authoring Phase**

A science teacher was recruited who updated a previously designed genetics course for secondary school pupils so that it met the requirements of the National Curriculum and the GCSE syllabus. The course consisted of text and graphics declarative material with some multimedia, simple exercises and a glossary. Material was changed whenever necessary to make it more appropriate to the age group and to make the information more accessible. New material was added and old material removed if considered no longer relevant to the syllabus under study. The course was divided into two parts. The first, Genetics1, was 48 pages long and covered the topics of inheritance, genes, variation and cell division. The second, Genetics2, was longer at 73 pages, and covered DNA structure, evolution and reproduction. The amount of information however was judged to be roughly the same for both courses.

The teacher then authored the two courses using the REDEEM tools to create ITSs with the following characteristics. 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. This was either done by creating a "core" section on a topic, such as genetic engineering and then adding additional material for a "hard" section or by creating two distinct sections such as "easy" and "hard" fertilization. The teacher created many questions and, for the majority of these, also authored multiple levels of hints. She described pages where she wanted students to reflect and developed a number of non-computer tasks (*e.g.* worksheets), which were authored to appear at appropriate times.

She then chose how to adapt the ITSs to the perceived needs of different learners. She created five different categories of learners based on her judgments of their relative aptitude (A to E). Each category was assigned different material and teaching strategies (summarized in Table 1). Some aspects of her strategy such as using no pre-tests, when to assign non-computer based tasks and whether to teach material from general to specific were common to all strategies.

Table 1. Summary of ITSs Created for Five Categories of Learner for Genetics1 & Genetics2

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

### CBT courses

Two CBT courses were constructed from the underlying courseware. Both courses included 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. Out of a possible 73 pages from Genetics2, 44 were included in CBT Genetics2. The order of presentation of the pages was the teacher's decision 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 by REDEEM and some blank pages to make notes. Thus, 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 only, i.e. half received REDEEM Genetics1 and

CBT Genetics2 and half CBT Genetics1 and REDEEM Genetics2 (see Figure 4). Subjects' scores at pre-test were used to ensure that there was an equal distribution of ability across the two conditions.

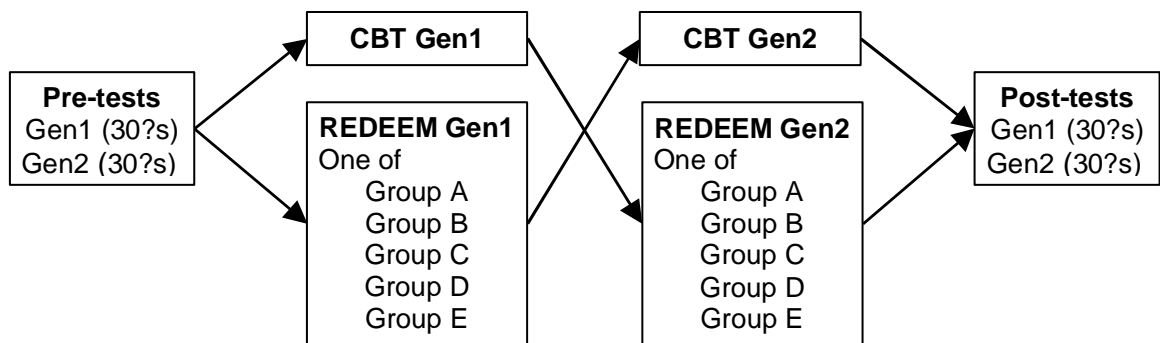


Figure 4. Design of Study

### Participants

Eighty-six mixed ability pupils from a state secondary school took part in the experiment. They were between 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 hence numbers in the five categories were 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, we decided the best solution was to use questions on the pre and post-tests that were based on material presented to all subjects (which may have a concomitant effect of removing the opportunities for high-flyers to shine). Furthermore, some of the material is directly questioned for (some) groups by REDEEM. Hence, performance on these items could be enhanced by simple memorisation rather than understanding. Rather than excluding the material, we chose to address this issue by creating three types of question:

- REDEEM questions - which were asked to all participants during their REDEEM intervention session;
- Surface Transformation questions – which addressed the same issue as a REDEEM question but had been manipulated so that simple memorization would not help (*e.g.* questions about 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 multi-choice quiz was developed. It consisted of 30 questions on Genetics1 and 30 on Genetics2 each further subdivided into 10 REDEEM, 10 Surface Transformation 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 the other half answered questions on Genetics2 followed by questions on Genetics1. Pre- and post- tests were the same.

### Procedure

1. Pre-tests were given to the participants in their school classroom just prior to the intervention
2. Intervention: The pupils came to the University of Nottingham to study the Genetics material. Each session lasted between 30 and 90 minutes. The minimum number of sessions a pupil attended for was two and the maximum was five. Variation in the number of sessions taken by different participants to complete the study was generally due to a difference in the quantity of note taking. 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. 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). Twelve subjects were excluded from the analysis due to non completion of pre or post-test.

Table 2. Pre and Post Test Scores (out of 30) by Course and Type of Environment

	REDEEM				CBT			
	Genetics1		Genetics2		Genetics1		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 ( $F_{1,72} = 74.52$ ,  $MSE = 7.62$ ,  $p < 0.0005$ ), that is, as predicted, post-test totals were greater than pre-test totals. There was also a significant main effect of course ( $F_{1,72} = 5.06$ ,  $MSE = 8.25$ ,  $p = 0.028$ ) with subjects scoring higher on Genetics2 than on Genetics1. There were no significant interactions. (Figure 5).

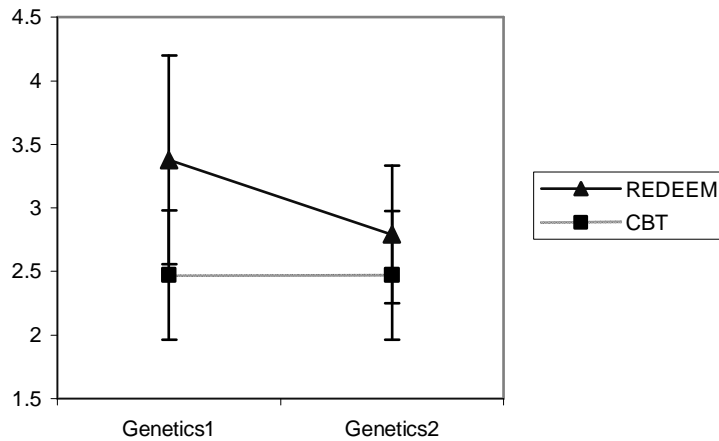


Figure 5. Improvement Scores by Type of Environment and Course

There was no correlation between learners' improvement on Genetics1 and Genetics2 ( $r = 0.12$ ) and so learners who made the greatest improvement on their REDEEM course were not the same ones who made greatest improvement of their CBT course.

An analysis of the effect of question type (REDEEM, Surface Transformation (ST) and Non-REDEEM (Non)) was performed to reveal if they were associated with same degree of improvement. We predicted that for the questions on the course they had experienced through REDEEM, they would perform significantly better on REDEEM and ST questions than on Non questions (Table 3). There should be no difference for CBT material as they were not asked any questions during that part of the intervention (Table 4). If significant differences are present at post-test, this would suggest that the differences we are observing are related to the actual questions rather than to the way the material has been presented to the subjects. 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.

Table 3. Pre and Post Test Scores (out of 10) by Question Type, Course and Time (REDEEM Only)

	Genetics1 (n = 40)						Genetics2 (n = 34)					
	Question type						Question type					
	RED		ST		Non		RED		ST		Non	
	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD
Pre-test	3.70	1.60	3.60	1.55	3.70	1.70	4.21	2.06	4.71	1.96	3.76	1.23
Post-test	5.23	1.97	4.78	2.06	4.38	1.69	5.62	2.31	5.74	1.90	4.12	1.79

Table 4. Pre and Post Test Scores (out of 10) by Question Type, Course and Time (CBT Only)

	Genetics1 (n = 34)						Genetics2 (n = 40)					
	Question type						Question type					
	RED/10		ST/10		Non/10		RED/10		ST/10		Non/10	
	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD
Pre-test	3.74	1.38	4.00	1.71	3.97	1.53	3.65	1.78	4.22	1.78	3.95	1.50
Post-test	4.56	1.74	4.76	1.91	4.85	1.60	4.75	1.92	5.15	1.70	4.40	1.46

For the REDEEM data, as before there was a significant main effect of time ( $F_{1,72} = 41.25$ ,  $MSE = 2.83$ ,  $p < 0.0005$ ) and question type ( $F_{2,144} = 9.99$ ,  $MSE = 2.45$ ,  $p < 0.0005$ ). There was a significant interaction between question type and course ( $F_{2,72} = 4.79$ ,  $MSE = 2.78$ ,  $p < 0.02$ ). Simple Main Effects showed that question type had a significant impact on Genetics2 ( $F_{2,144} = 12.40$ ,  $p < 0.001$ ) but not on Genetics1 ( $F_{2,144} = 1.58$ ). Furthermore, there was a significant interaction between time and question type ( $F_{2,144} = 5.11$ ,  $MSE = 1.67$ ,  $p < 0.007$ ). Simple Main effects analysis revealed that there were no significant 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 Surface Transformation 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 ( $q = 2.70$ ). However, the analysis of the CBT data showed a single significant effect, that of time ( $F_{1,72} = 43.71$ ,  $MSE = 1.71$ ,  $p < 0.0005$ ) (see Figure 6).

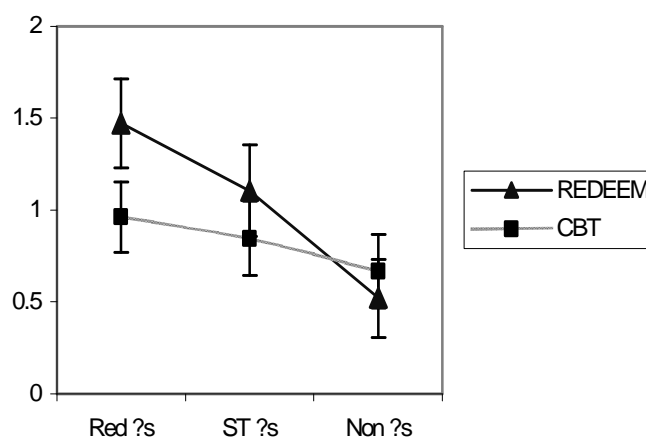


Figure 6. Improvement Scores by Question Type and Type of Environment (collapsed across course)

#### **Relation between Pre-test Scores, Teachers' Categorisation and Learning Outcomes**

The relation between pre-test scores and learning outcome was examined. Unsurprisingly, there was a significant positive correlation between pre-test scores and post-test scores ( $r = 0.70$ ,  $N = 74$ ,  $p <$

0.0005), but no significant relationship between pre-test scores and improvement scores, ( $r = -0.10$ ,  $N = 74$ ,  $p = ns$ ). This indicates that learners at all levels of prior knowledge made similar improvements from pre to post-test.

Table 5 shows the scores for the different teachers categories based again on their performance on the pre-tests. Learners' 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  $p < 0.001$ ). This suggests that the teacher aptitude categorisation, which was blind to the pre-test score, had been accurate. Tukey tests showed that 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;  $p < 0.0005$  in all cases).

Table 5. Pre-test Scores (out of 60) by Student Category

	Pre-test Scores	
	$\bar{x}$	S.D.
A (n=8)	33.62	5.26
B (n=30)	26.13	2.53
C (n=20)	20.55	1.64
D (n=10)	17.70	2.36
E (n = 6)	16.83	6.94

To determine if any learner category had differentially improved we examined pre and post-test performance. To achieve a sufficiently large number of subjects per cell we collapsed across cells to create two meta-groups of high and low scorers. We felt this grouping was justified by the difference in pre-test scores (see above) and the differences in the teacher's authoring (Table 1). Mean totals at pre- and post-test are shown in Table 6. To reduce the complexity of the analysis and because previous analyses had confirmed no interactions between course (Genetics1/Genetics2) 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 meta-group.

Table 6. Pre and Post Test Scores (out of 30) by Two Meta -Categories and Time of Environment

	High (n=38)				Low (n=36)			
	REDEEM		CBT		REDEEM		CBT	
	$\bar{x}$	S.D.	$\bar{x}$	S.D.	$\bar{x}$	S.D.	$\bar{x}$	S.D.
Pre-test	14	3	13.71	2.63	9.42	2.85	9.72	2.84
Post-test	17.14	4.34	16.11	3.15	12.5	3.95	12.28	3.2

Analysis by [2 by 2 by 2] ANOVAs on the subjects' REDEEM and CBT questionnaire data showed a significant main effect of time ( $F_{1,72} = 75.45$ ,  $MSE = 7.64$ ,  $p < 0.0001$ ) and meta-group ( $F_{1,72} = 63.76$ ,  $MSE = 21.02$ ,  $p < 0.0001$ ). None of the interactions were significant. Both groups of students improved similarly when learning with REDEEM and CBT. This is graphed as improvement scores in Figure 7.

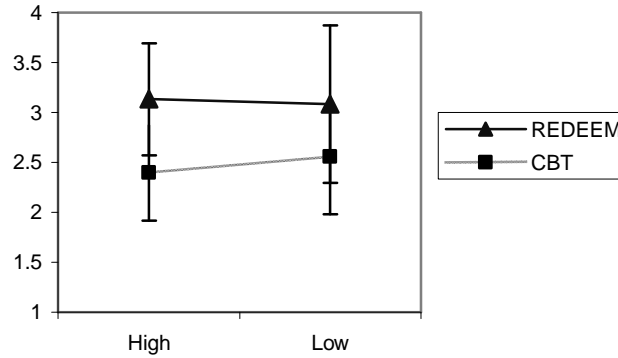


Figure 7. Improvement Scores by Meta-group by Type of Environment (collapsed across course)

### Process Measures for REDEEM and CBT

Analysis of the paper test data showed that overall students improved but the degree of improvement was not influenced by the nature of the learning environment or their categorization. Furthermore some students 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). Hence, we explored a number of measures of system use to determine if **how** students used the system influenced what they learnt.

Table 7. Time in Seconds (for each course and page) by Student Category and Course

		Genetics1			Genetics2		
		$\bar{x}$	No of pages	Time per Page	$\bar{x}$	No of pages	Time per Page
A (n=8)	RED.	9316	44	212	7862	60	131
	CBT	3180	29	110	3050	40	76
B (n=30)	RED.	5126	44	114	5195	50	104
	CBT	2874	29	99	2581	40	65
C (n=20)	RED.	4727	32	140	2960	44	67
	CBT	2912	29	100	3720	40	93
D (n=10)	RED.	2911	30	83	2773	44	63
	CBT	2943	29	101	3256	40	81
E (n=6)	RED.	2961	30	93	2870	44	65
	CBT	3199	29	110	1847	40	46

To calculate the amount of time the participants had spent learning with the packages, the time they spent over each session from their first to last mouse click was totalled. The time per course is obviously influenced by the fact that each course differed in length and the REDEEM courses also differed depending upon the student category. Hence, we divided the total amount of time by the number

of pages in the course to determine the mean amount of time in seconds 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 learners in Category A spent significantly longer per page than learners in Groups D and E ( $q = 4.02$ ,  $p < 0.05$  and  $q = 4.71$ ,  $p < 0.05$ ). There was a significant interaction between environment and course ( $F_{1,67} = 8.694$ ,  $MSE = 3197.90$ ,  $p < 0.004$ ). 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.

We inspected this data to see if learners 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 with increased time being positively associated with improvement ( $r = 0.29$ ,  $N = 74$ ,  $p < 0.015$ ) and post-test performance ( $r = 0.33$ ,  $N = 74$ ,  $p < 0.005$ )

### Use of Notes

Students were provided with a pen and paper for the CBT and an on-line notes tool. They were told that writing notes would help them understand and remember. To test if this statement was true we performed a simple analysis of their notes exploring only the amount of notes written rather than the quality of those notes.

Table 8. Number of Words Written by Student Category and Nature of Environment

	Words in Notes in REDEEM		Words in Notes in CBT	
	$\bar{x}$	S.D.	$\bar{x}$	S.D.
All Subjects (72*)	240.12	198.97	98.79	143.73
A (n= 8)	384.63	291.07	78.63	94.03
B (n =28*)	226.93	183.05	87.43	140.88
C (n= 20)	243.50	214.67	91.65	152.01
D (n= 10)	190.40	91.96	95.70	101.33
E (n= 6)	180.67	169.27	207.67	227.32

- Two students in the study 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 than when learning with CBT ( $F_{1,64} = 17.65$ ,  $MSE = 26246$ ,  $p < 0.001$ ). There were no further effects.

Table 9. Correlations between Notes Written and Performance by Student Category and Environment

		Pre-test	Post-test	Improvement	Improvement pre-test factored out
All subjects	RED.	0.22 (p=0.064)	0.30 (p=0.010)	0.15 (p=0.208)	0.22 (p=0.061)
	CBT	-0.09 (p=0.468)	-0.01 (p=0.947)	0.08 (p=0.489)	0.054 (p=0.653)
A (n= 8)	RED.	0.14 (p=0.735)	-0.02 (p=0.640)	-0.49 (p=0.222)	-0.47 (p=0.285)
	CBT	-0.33 (p=0.430)	-0.53 (p=0.179)	-0.19 (p=0.653)	-0.47 (p=0.284)
B (n= 28)	RED.	0.13 (p=0.525)	0.35 (p=0.067)	0.28 (p=0.151)	0.33 (p=0.09)
	CBT	-0.05 (p=0.789)	-0.19 (p=0.326)	-0.15 (p=0.451)	-0.19 (p=0.354)
C (n= 20)	RED.	0.09 (p=0.701)	0.16 (p=0.494)	0.07 (p=0.761)	0.18 (p=0.452)
	CBT	-0.34 (p=0.149)	0.01 (p=0.968)	0.24 (p=0.299)	0.08 (p=0.734)
D (n= 10)	RED.	-0.28 (p=0.429)	0.62 (p=0.058)	0.76 (p=0.011)	0.73 (p=0.025)
	CBT	0.44 (p=0.207)	0.55 (p=0.104)	0.21 (p=0.562)	0.56 (p=0.120)
E (n=6)	RED.	0.01 (p=0.979)	-0.44 (p=0.381)	-0.18 (p=0.735)	-0.44 (p=0.454)
	CBT	0.62 (p=0.193)	0.74 (p=0.091)	0.13 (p=0.801)	0.55 (p=0.335)

To explore if there was a relationship between note taking and performance, Pearson's correlations were carried out on the number of notes and performance (pre-test, post-test, improvement) combined for all subjects and for each ability group separately. Table 9 shows that there was a weak relationship between notes taking and learning outcomes for the REDEEM group. Students who wrote more notes performed better on post-test in the REDEEM condition ( $r = 0.30$ ,  $N=72$ ,  $p<0.01$ ) but not in the CBT condition ( $r= -0.01$   $N = 72$ ,  $p=ns$ ).

### REDEEM only process measures

REDEEM automatically captures information about an individual's use of system features as it uses as the basis of student reports. In addition to the information about time spent reading information or responding to questions, it also automatically logs number of attempts at questions, questions scores, and the number of hints either provided or requested. Analysis of these measures allows us to explore firstly how these system features were used and secondly, whether was a systematic relationship between behaviour with REDEEM and performance. We present these analyses by student category and have for simplicity collapsed across course. Hence data is presented for the learner's REDEEM interaction irrespective of whether it was Genetics1 or Genetics2.

### Question answering

To explore if there was a systematic relationship between the learners' performance on questions during their intervention session and their incoming knowledge or post-intervention performance, we analysed their question performance. Questions were classified into those right first time and everything else which could include answers right on a second, third attempt or never answer correctly.

**Table 10. Percentage of Questions Right First Time by Student Category**

	% ?s right 1 <sup>st</sup> time	
	$\bar{x}$	S.D.
All Ss (72)	47.06	14.60
A (n=8)	60.85	14.53
B (n=28)	47.31	12.87
C (n=20)	48.97	14.60
D (n=10)	37.44	11.47
E (n=6)	37.25	13.14

A [5 by 1] ANOVA on percentage of questions right first time confirmed the influence of category ( $F_{4,67} = 4.31$ ,  $MSE = 179.7$ ,  $p < 0.004$ ) with learners in Category A performing better than those in D or E ( $q = 23.41$   $p < 0.004$  and  $p = 23.61$ ,  $p < 0.015$ ).

**Table 11. Correlation between Questions Right First Time and Performance by Student Category**

	Pre-test	Post-test	Improvement	Post-test (pre-test factored out)
All (n=72)	+0.33 ( $p < 0.004$ )	+0.63 ( $p < 0.0005$ )	+0.43 ( $p < 0.0005$ )	+0.57 ( $p < 0.0005$ )
A (n=8)	+0.72 ( $p < 0.04$ )	+0.50 ( $p = 0.207$ )	-0.34 ( $p = 0.405$ )	-0.10 ( $p = 0.825$ )
B (n=28)	-0.07 ( $p = 0.739$ )	+0.33 ( $p < 0.080$ )	+0.36 ( $p < 0.061$ )	+0.36 ( $p = 0.068$ )
C (n=20)	-0.01 ( $p = 0.958$ )	+0.75 ( $p < 0.0005$ )	+0.57 ( $p < 0.009$ )	+0.76 ( $p < 0.0005$ )
D (n=10)	-0.04 ( $p = 0.923$ )	+0.82 ( $p < 0.004$ )	+0.82 ( $p < 0.004$ )	+0.85 ( $p = 0.004$ )
E (n = 6)	0.04 ( $p = 0.936$ )	+0.73 ( $p < 0.097$ )	+0.23 ( $p = 0.656$ )	+0.73 ( $p = 0.159$ )

There was a consistent relationship between question behaviour and learners' prior knowledge. Those students who scored higher on pre-test were more likely to answer the question right first time whereas those who scored lower were more likely to fail to answer correctly on their initial attempt. This pattern of result was confirmed when we examining learning outcomes. Students with higher post-test scores performed better on the system whereas those with lower scores again were less likely to get the questions right initially during the intervention. If we control for pre-test scores, the significant relationship between answering questions correctly and improvement scores remains ( $r = 0.57$ ,  $N = 72$ ,  $p < 0.0005$ ). Students who knew more to begin with, answer REDEEM questions correctly and that answering REDEEM questions correctly is associated with increased learning outcomes.

#### 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 (which only shows a hint

when a learner gets the answer wrong) and for groups D & E, hint on error and request (which in addition, allows students to request hints at any time). Secondly, for groups D & E then the teacher chose that students should have only two attempts before being told the right answer. Consequently, if they chose not to ask for hints, these students were limited to only one hint before seeing the right answer.

Table 12. 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 12 shows that there was a low uptake of hints on request. This is not because students knew the answers to these questions as Table 10 shows that students in these categories were only likely to get the question right first time around a 1/3<sup>rd</sup> 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. Overall, students did not really use this feature.

Analysis by two [5 by 1] ANOVA revealed that total number of hints was related to student category ( $F_{4,67} = 17.01$ ,  $MSE = 56.02$ ,  $p < 0.001$ ) (Table 13). Category B learners received significantly more hints than all other categories and there were no other differences ( $q=12.82$ ,  $p<0.001$ ;  $q=17.17$ ,  $p<0.001$ ;  $q = 9.47$ ,  $p<0.01$ ;  $q= 11.57$ ,  $p<0.001$  for B with A,C,D and E respectively). The percentage of questions with hints displayed also showed differences between groups ( $F_{4,67} = 4.31$ ,  $MSE = 179.7$ ,  $p < 0.004$ ). This time Category A learners had less questions with help than most other groups ( $q=15.00$ ,  $p<0.03$ ;  $q=25.54$ ,  $p<0.001$ ;  $q = 23.78$ ,  $p<0.01$  for B,D and E). This dissociation in measures is partly explained by way that help (on error) is determined by the number of questions a learner gets wrong and the number of attempts that they are allowed at each question. The students in category A got fewer questions wrong, hence the low percentage of questions with help. However, they had many attempts to get the question right and received a hint on each incorrect attempt and so the relatively high number of total hints.

Table 13. Number of Hints Shown on Error by Student Category

	No of Qs	Total no of hints on error		% Qs with hints on error	
		$\bar{x}$	S.D.	$\bar{x}$	S.D.
All Ss (n=72)		18.64	10.41	47.09	13.93
A (n= 8)	28 or 26	15.25	7.76	31.90	12.58
B (n= 28)	30 or 28	28.07	10.07	46.90	11.23
C (n= 20)	23 or 25	10.70	3.31	45.67	14.01
D (n= 10)	23 or 25	13.90	2.69	57.44	12.62
E (n= 6)	23 or 25	13.50	2.35	55.68	11.81

To explore if there was a relationship between use of help and performance, Pearson's correlations were carried out between help (total number of hints, hints on error and hints on request) and performance (pre-test, post-test, improvement) combined for all subjects and for each ability group separately (Table 14).

Table 14. Correlations between Amount of Help and Performance by Student Category

		Pre-test	Post-test	Improvement	Post-test (pre-test factored out)
All subjects (n = 72)	Total hints	0.12 (p=0.335)	-0.16 (p=0.180)	-0.29 (p=0.014)	-0.27 (p=0.023)
	Hint on error	0.16 (p=0.191)	-0.13 (p=0.289)	-0.29 (p=0.014)	-0.26 (p=0.029)
A (n= 8)	Hint on error	-0.83 (p=0.01)	-0.67 (p=0.07)	0.27 (p=0.525)	-0.10 (p=0.826)
B (n= 28)	Hint on error	-0.11 (p=0.576)	-0.57 (p=0.002)	-0.50 (p=0.007)	-0.56 (p=0.002)
C (n= 20)	Hint on error	0.038 (p=0.874)	-0.78 (p<0.0005)	-0.60 (p=0.005)	-0.79 (p<0.0005)
D (n= 10)	Total hints	0.50 (p=0.146)	0.03 (p=0.938)	-0.25 (p=0.491)	-0.11 (p=0.788)
	Hint on error	-0.02 (p=0.952)	-0.87 (p=0.001)	-0.83 (p=0.003)	-0.89 (p=0.001)
	Hint on request	0.43 (p=0.211)	0.37 (p=0.297)	0.12 (p=0.750)	0.30 (p=0.430)
E (n= 6)	Total hint	0.13 (p=0.813)	0.19 (p=0.721)	-0.05 (p=0.926)	0.18 (p=0.770)
	Hint on error	-0.06 (p=0.908)	-0.44 (p=0.386)	-0.11 (p=0.843)	-0.43 (p=0.465)
	Hint on request	0.18 (p=0.730)	0.51 (p=0.305)	0.02 (p=0.976)	0.50 (p=0.387)

A cursory inspection of the data would suggest that reading hints make students worse (*e.g.* for improvement scores and total hints,  $r = 0.29$ ,  $N = 72$ ,  $p < 0.014$ ). However, this conclusion is not warranted. As students receive hints in most categories only when they make an error, what this figure is at least partially obscuring is the number of attempts students need to answer the question correctly. Once this has been partialled out the significant relationship between total hints and improvement disappears ( $r = -.21$ ,  $p = ns$ ).

## DISCUSSION

### Learning Outcomes

The results of this study showed that whilst pupils 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. REDEEM improves learning by 0.2 sigmas compared to CBT. This was true for learners in all student categories. Furthermore, the degree to pre-test to post-test improvement was significantly significant but not as substantial as we would have liked. Learners' 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 learners improved more on questions (and surface transformations of questions) that REDEEM had presented during their intervention. For the CBT courses, there was no difference in the three types of question on post-test but for the REDEEM ITSs there was significant improvement from pre-test to post-test only on the REDEEM and Surface Transformation questions.

We also examined learners' performance according to their assigned student category. It was evident that the teacher had good knowledge of her 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

We predicted that REDEEM should slow learners down simply because more learning activities are supported (answering questions, reading feedback, prompts to write notes). This is what we observed. Students using REDEEM spent an average of 17 more seconds a page than students interacting with CBT. Furthermore, whilst time did not correlate with performance for CBT courses, it did for REDEEM courses with increased time being associated with higher post-test and improvements scores. It seems that time on course *per se* is not an important determiner of learning outcomes, but that if students spend time on the additional activities provided by REDEEM then this may improve their chance to learn.

To gain some insight into how students' note taking influenced learning we performed a rudimentary analysis of their online and paper notebooks, simply counted the number of words written without examining either the accuracy or quality of the statements. Essentially, we viewed the amount of notes learners made as indicative of the amount of effort they were prepared to expend on learning. Students made many more notes using REDEEM than they did when learning with the CBT. This may either be because they preferred to type their answers and/or because REDEEM prompts learners to write notes at places teachers have indicated as "reflection points". 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 learners to write notes, and furthermore, writing notes is a reasonable predictor of learning outcomes. One explanation that is consistent with this result is that the REDEEM prompts were providing learners 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 learners 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 the material as CBT.

### **REDEEM Only Process Measures**

A number of process measures that describe students' interactions with REDEEM are recorded as part of the student history that REDEEM offers to teachers. These are particularly useful way to explore the behaviour of learners in different student categories. Apart from the time data discussed above, the first variable of interest is the number of attempts students need to get a question correct. Unsurprisingly, this showed a significant relationship with pre-test, with students who knew more before learning with REDEEM getting more answers right first time. However, there was a positive significant relationship 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. 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 got 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 at differentially easier for these learners even though they were given only 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. Depending on the authors' decisions, this can either be on error or in addition can be provided when a student requests it. Overall, there was no evidence of the positive or negative impact of help on learning outcomes. There was a significant relationship between help received on error and improvement but this is explained by the answer being wrong not by the help received. Students in Group B received significantly more hints than all other categories. This would suggest that these students (who also received only medium and hard questions) found these questions fairly difficult. As they did not have more questions with hints than other categories, it would also suggest they required multiple attempts to get the question right.

Two categories of learners (D & E) were able to ask for help on request. They rarely took advantage of this feature, requesting hints on only around 10% of questions even though their first answer to this question was wrong around 62% of the time. Given this low uptake and the skewed distribution (one learner accounted for 35% of all requested hints), unsurprisingly there was no relationship with performance measures. What we do not know is whether learners did not find the hints useful, were unconcerned about their performance, did not realise they needed help, or preferred to try and find out the answer for themselves. However, the low 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 feature. Student 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 factored out, many of these factors remain significant.

Overall, the results of the study did not find that REDEEM ITSs were any more effective than the CBT that they were based on. We will discuss the reasons for this finding at length in the general discussion. However, one plausible explanation that we sought to rule out was that the unnatural situation of using university laboratories instead of school classrooms had reduced the validity of the experiment. We had been unable to use school computing facilities because of timetable clashes but bringing pupils into the University led to a series of less than ideal circumstances. 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, pupils 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 significantly 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 ITSs. He started from the authoring from the previous experiment and maintained many of the questions. However, he substantially changed

the structure of the material (see Ainsworth, Clarke & Gaizauskas, 2002), rewrote many of the hints and recited reflection points. This teacher chose a coarser-grained description of learners than in Study One. Three different categories of learners were created that corresponded to different sets (classes differentiate 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 three ITSs created with REDEEM did include some different material and questions. However, unlike Study One, the teacher choose to give all groups the same teaching strategy (summarized in Table 15).

Table 15. Summary of ITSs Created for Three Categories of Learner for Genetics1 & Genetics2

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

### **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 pupils from a local City Technology College were originally selected to take part in the experiment. These students were in three differences classes grouped by ability. Unfortunately, a very substantial number of students were not able to complete the whole experiment. 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 two weeks into the intervention. The supply teachers who replaced him were unfamiliar with the material, experiment and students. These factors in addition to the standard problems of absences meant that we only had full data from 15 pupils. They were between 14 and 15 years old and there were 9 boys and 6 girls. Of the 15 pupils, 11 were in the top ability set, and 4 were in a lower ability set.

### Materials

Pre and post-tests were similar to Study One, but with new questions to replace material not covered in these version of the CBT/ITSs. Thus, a 60 item multi-choice quiz was used (30 questions on Genetics1 and 30 on Genetics2) further subdivided into 20 REDEEM, 20 Surface Transformation and 20 Non-REDEEM questions.

### 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, either in one of the school's computing labs, or in a classroom set up with laptops. Each session lasted either 45 or 90 minutes, depending on whether it was a single or double lesson, and there was 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 environments (REDEEMGenetics1/CBTGenetics2, REDEEMGenetics2/CBTGenetics1).

Table 16. 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 ( $F_{1,13} = 54.39$ ,  $MSE = 3.48$ ,  $p < 0.0005$ ) with post-test scores higher than pre-test scores. There was no significant main effect of course nor environment but the interaction between time, course and environment approached significance ( $F = 4.64$ ,  $MSE = 4.43$ ,  $p = 0.051$ ). Simple Main effects analysis showed that subjects' scores on Genetics1 and Genetics2 improved whether they score the course under REDEEM or as CBT, accept 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  $F_{1,13} = 3.04$ ,  $MSE = 3.48$ ,  $p = ns$ ). Figure 8 graphs this interaction as improvement in performance (*e.g.* the scores for REDEEM GENETICS1 at post-test were 5.5 higher at post-test).

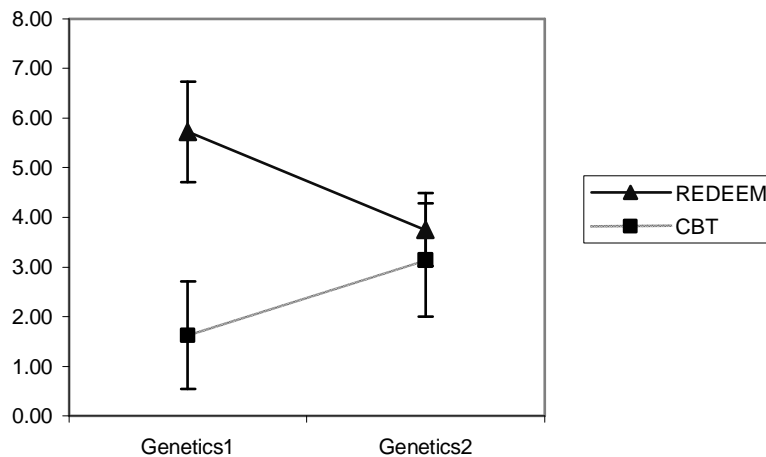


Figure 8. Improvement Scores by Type of Environment and Course

To examine if learners 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 ( $r = 0.883$ ,  $N = 15$ ,  $p < 0.0005$ ), but no significant relationship between pre-test scores and improvement scores, ( $r = -0.302$ ,  $N = 15$ ,  $p = ns$ ). This indicates that learners at all levels of prior knowledge made similar improvements from pre to post-test. To examine if there was an individual differences effect (*i.e.* that certain students learnt more irrespective of condition), learners' improvements scores on REDEEM and CBT were correlated. There was no relationship ( $r = -0.17$ ) between learners' improvement on the two courses.

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.

**Table 17. Pre and Post Test Scores (out of 10) by Question Type, Course and Time (REDEEM Only)**

	Genetics1 (n = 7)						Genetics2 (n = 8)					
	Question type						Question type					
	RED		ST		Non		RED		ST		Non	
	$\bar{x}$	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

**Table 18. Pre and Post Test Scores (out of 10) by Question Type, Course and Time (CBT Only)**

	Genetics1 (n = 8)						Genetics2 (n = 7)					
	Question type						Question type					
	RED/10		ST/10		Non/10		RED/10		ST/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 ( $F_{1,13} = 37.18$ ,  $MSE = 1.50$ ,  $p < 0.0005$ ) and question type ( $F_{2,26} = 6.70$ ,  $MSE = 1.77$ ,  $p = 0.005$ ) with both REDEEM and Surface transformation questions being scoring significantly better than non-redeem questions ( $q = 4.4$ ,  $p < 0.05$  and  $q = 4.53$ ,  $p < 0.01$  respectively). 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 ( $F_{1,13} = 12.44$ ,  $MSE = 1.14$ ,  $p = 0.004$ ). The interaction between question type and course was also significant ( $F = 5.37$ ,  $MSE = 11.24$ ,  $p = 0.011$ ). 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.

### Process Measures

The time on task data from school classroom is too noisy to be sensibly interpreted with time on task information 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 were provided with a pen and paper for the CBT and an on-line notes tool. They were told that writing notes would help them understand and remember. To test if this statement was true we performed a simple analysis of their notes exploring only the amount of notes written rather than the quality of those 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 (Table 19).

**Table 19. Correlations between Notes Written and Performance by Student Category and Environment**

		Pre-test	Post-test	Improvement	Post-test (pre-test factored out)
Subjects	No of Notes				
(n = 15)	$\bar{x} = 125.20$ S.D. = 171.91	-0.07 (p=0.82)	-0.03 (p=0.92)	0.16 (p=0.58)	0.14 (p=0.62)

Given that students did not take any notes at all when using the CBT, obviously REDEEM promoted note taking. However, there was no systematic relationship between number of words written and any aspect of students' performance

### REDEEM Process Measures

#### Question Answering

In order to explore if there was a systematic relationship between the students' performance on questions during their intervention session and their incoming knowledge or post-intervention performance, we analysed their question performance.

**Table 20. Correlation between Percentage of Questions Right First Time and Performance**

		Pre-test	Post-test	Improvement	Post-test (pre-test factored out)
Subjects	%s right 1 <sup>st</sup> time	+0.67	+0.80	+0.13	+0.58
(n= 15)	$\bar{x} = 56.96\%$ S.D. = 14.59	(p=0.006)	(p<0.001)	(p=0.64)	(p=0.03)

As can be seen from Table 20, there was a consistent relationship between question behaviour and students' prior knowledge. Those students who scored higher on pre-test were more likely to answer the question right first time whereas those who scored lower were more likely to fail to answer correctly on their initial attempt. This pattern of result was confirmed when we examining learning outcomes. Students with higher post-test scores performed better on the system whereas those with lower scores were less likely to get the questions right initially. The association for between post-test and answering the question correctly remained even when pre-test scores were partialled out.

## Use of Help

Table 21. Number of Hints Provided on Error

	No of Qs	No of hints on error	% Qs with hints on error		
Subjects	$\bar{x}$	$\bar{x}$	S.D.	$\bar{x}$	S.D.
(n= 15)	32.80	20.00	8.74	32.43	12.34

To explore if there was a relationship between use of help and performance, Pearson's correlations were carried out between help (total number of hints, hints on error and hints on request) and performance (pre-test, post-test, improvement, improvement with pre-test partialled out) combined all subjects (Table 22).

Table 22. Correlations between Amount of Help and Performance

		Pre-test	Post-test	Improvement	Post-test (pre-test factored out)
Subjects	No of hints	-0.67	-0.77	-0.09	-0.52
(n = 15)	provided	(p<0.006)	(p<0.001)	(p= ns)	(p=0.054)

Given that the only time hints were available to students, these figures again simply reflect the number of times students get questions wrong as when this is partialled out there significant relationships with pre and post-test disappear.

## DISCUSSION

### Learning Outcomes

Students performance at post-test was significantly higher than it was at pre-test, but this main effect was modified by the predicted three way interaction, i.e. that pupils scores would improve more 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. However, contrary to our prediction this was not true for Genetics2. Students moving from REDEEM to CBT to 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. Although the REDEEM data showed the predicted effects of REDEEM and Surface Transformation questions being answered better, unfortunately this effect is not modified by time. It would appear that by chance the students were less familiar with the Non-Redeem questions than with the other types of question. 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 nature of environment on post-test performance.

### **Process Measures**

Learners in the CBT condition did not use their pen and paper books to write notes. These groups of learners were particularly computer-literate and did not expect to have to use a pen when they could use a keyboard! Unsurprisingly REDEEM students wrote more notes (although only an average of 125 per person). Perhaps because of this low value, there was no correlation between amount of notes and learning outcomes. There was a strong association between the percentage of questions answered correctly first time and performance. Students with higher post-test scores get more answers right first time. There was a positive significant relationship with post-test scores, which remained even when pre-test scores were partialled out. Again, 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. Any association between learning outcomes and help provision will also be explained by this measure as the author chose to only provide help on error.

### **GENERAL DISCUSSION**

Learners 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, there was a consistent indication of REDEEM's influence on learning outcomes although the scale of this advantage varied greatly. For Study One, REDEEM scores only improved by an average 0.64 more questions 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 5.7 questions whereas CBT improved by 1.6 questions. 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 were also an indication of potential benefit in that 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

that given the generally inappropriately low times spent with the learning environments, that REDEEM appeared to slow the students down.

However, this is obviously not quite the strong endorsement of REDEEM's benefit that we were looking for. Although REDEEM tended to be better than CBT in all four situations, in only one case was this difference significant. In order to understand these results, we face a very large credit assignment problem. Evaluating the effectiveness of an ITS authoring environment requires consideration of many different interacting factors. Firstly, authoring tools normally offer teachers a constrained set of options to make ITS construction faster. In fact, REDEEM is one of the most rigid of all ITSATs as interaction is primarily via pre-defined dimensional ratings. Secondly, the ITS shell must then deliver the decisions in an effective way. Thirdly, teachers' authoring must be sensitive to learners' needs. In REDEEM's case, this primarily concerns the interactivity features (*e.g.* informative questions, supportive feedback on answers, helpful hints, reflection points at appropriate places) and course structure (an appropriate categorisation and sequence of material). It also means that differentiation should be appropriate (*i.e.* suitable material, questions and teaching strategy). Finally, a ITS delivered with REDEEM also depends on external underlying courseware. Accordingly, a REDEEM ITS is a combination of the options for authoring offered to users, the authors' decisions, and the systems' interpretation and delivery of those decisions using the courseware. Consequently, we have tried to determine which of these factors may have led to these patterns of result.

Both teachers had detailed knowledge of the topic, which allowed them to create a very 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 comes from that fact that the second teacher changed few of questions, reflection points and non-computer-based tasks that the first author created. Although some of the hints were rewritten and the course was reorganized when he created ITSs for the second study. We know from prior studies that teachers will substantially change ITSs if they are unhappy with them (Ainsworth *et al.*, 1999), so we believe that on the whole both teachers were happy with each other's authoring.

The basis of the teachers' classification of their pupils into different categories also seemed fairly unproblematic. In Study One, the teacher also 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, categorisation was not determined by the teacher but by a previously agreed streaming procedure which was monitored by the school.

There may be more grounds for concern about how these categories interacted with REDEEM's differentiation features. In Study One, student categories were assigned really quite different material (see Ainsworth, Clarke & Gaizauskas, 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 in the categories had little chance of understanding but that he needed to include it because it was on 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 fixing fairly strict number of attempts at questions for some groups. In Study Two, we would have tended to 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 learners 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 much of the results. In Study One, all groups of learners improved similarly from pre to post-test and each group received a different teaching strategy. The single most consistent change that we would have made would have been to allow help on request (only two groups of students in Study One had this feature and none in Study Two). However, use of this feature was disappointing low and so it would have been unlikely to have made a significant difference. A more fundamental question that we will come back to later is whether the different teaching strategies that REDEEM allows could ever impact significantly on learning.

The fourth type of explanation concerns the courseware. Unlike other ITS authoring tools, REDEEM relies on the pre-existing content. REDEEM's goal is to enhance the teaching of this content by providing additional interactivity and to differentiate this material and interactivity by providing alternative teaching strategies and content. If this underlying courseware is not of good quality then REDEEM may be unable to do much to enhance it. Furthermore, if the courseware is already rich in interactivity and allows for learners with different needs then perhaps REDEEM's feature are superfluous. This is not an explanation that appears likely in this case. The CBT had already been used in a school classroom and contains much clearly presented and relevant information. However, given its very limited interactivity and scope for different interpretations of the material, it could still be enhanced by REDEEM's features.

The four interrelated factors we identified were a) authoring tools functions, b) ITS shell functions, c) courseware, and d) authoring decisions. Analysis of the authoring and use of REDEEM in these studies make explanations based on courseware and authoring unlikely. We could see no reason why REDEEM Genetics1 as authored by the second teacher (which was significantly better than the CBT) was substantially different to either his REDEEM Genetics2 course or to the other teachers' authoring. It possible that had REDEEM been more flexible and adaptive that the significant advantage observed for REDEEM Genetics1 in Study 2 would have been greater and replicated in the other conditions.

There are two other factors that are worth considering when investigating why REDEEM appeared to be beneficial in some circumstances and not in others, namely the wider context of the studies and the

learners. The two studies were run with students who had a genuine need to learn the material as part of their education and the material they studied with both environments had been deemed suitable and relevant by their class teachers. As the authors of the ITSs were their class teachers, they had detailed knowledge of both the domain and the students. However, in Study One, pupils 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, students being restreamed into other classes seriously compromised that 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 very time-consuming and had to be 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 design 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 more authentic evaluations in actual classrooms provide much less control over the setting and lead to subsequent difficulty 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 learners in these studies. One reason why difference between the results for CBT and REDEEM may be lessened is that individual learners 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, learners 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 learners' 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 very notable that many pupils 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, the results for Study Two, which found that REDEEM's increased learning relative to the CBT for Genetics1 excluded the learners 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 learners who took advantages of REDEEM's features learnt more than those who did not.

### **Exploring the Difference between the REDEEM ITSs and the CBT**

The results of these two studies showed that the REDEEM ITS authoring environment in the hands of a knowledgeable teachers can, in certain circumstances, lead to increased learning outcomes relative to CBT. However, this effect is not as robust as we would like. In three cases, learners interacting with REDEEM ITSs learnt only a small amount more than those learning with CBT and in only one case, did students being taught with REDEEM learn significantly more than those given CBT. Inspection of the process data suggested that REDEEM would only improve learning for those students that chose to engage with the additional interaction features that it provides. This is not that unexpected as the REDEEM only differs from the underlying courseware in three ways: 1) the structure of material, 2) the interactive features, and 3) the macro-adaptation features. Unlike traditional ITSs, REDEEM does not micro-adapt to learners. In these studies, the authors created a structure for the CBT that while it differed to the REDEEM ITSs (which also differed to each other) was adapted to their view of teaching. As a consequence, we reduced the difference between traditional use of CBT and REDEEM and hence the likelihood of differences in learning outcomes. Therefore the main differences between the two environments lie in the interactive and 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 would suggest that in these studies 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 ITSs. Secondly, the process measures that identified the learners that had made the greatest improvement showed that it was those learners who interacted most with the environment. The fact that degree of interaction with the environment predicted learning outcomes does not seem contentious but the question that remains is why we did not observe benefits from the macro-adaptation. We propose three potential reasons; 1) Macro-adaptation is unimportant, 2) That the macro-adaptation in this study was inappropriate and 3) That the macro-adaptation was potentially was 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 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) However, one reason why adapting the teaching strategy to different categories of learners was not associated with obviously better performance was that authors' made inappropriate decisions in the nature of the categorisation or in how content or strategy was adapted to these categories.

There is little doubt that the teachers' categorisation of learners, which was based on their perceived aptitude, was accurate. However, it could be the aptitude was not the most appropriate dimension to use to rank the students or that it could have been supplemented by other factors. For example, some students in Group A and B's who had high control over when they answered questions did not use the feature appropriately. Some chose not to answer the questions until they that the program 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, we could imagine a situation where factors such as motivation and self regulatory skills could be used to set teaching dimensions concerning issues 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.

Secondly, assuming the classification was appropriate and accurate, did teachers make the best decisions about how to assign content and strategies? From inspection of the ITSs, 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 (i.e. fell within their Zone of Proximal Development). 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 the aptitude, there should have been equal performance across all categories. However, we believe that questions right first time indicates not just prior knowledge and learning, but also how 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 of more learner control in higher groups which is

in line with research, which tends to find that those students who score higher on pre-tests learn better with high control (*e.g.* for a review see Williams, 1996). She also provided higher ability learners 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 in rule-given environments. However, there may be more disagreement between the author's decision and previous findings with help seeking as in this case 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. Overall, there is no real evidence to suggest 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 weak and they may have little impact in the classroom. The number of students in each category in Experiment One would often not have been sufficient to allow identification of benefits unless they were very substantial. Moreover, from this design they 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, a 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, learning outcomes and processes of students' interaction with simple ITSs created with REDEEM 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 (effect size of 1.33), then there would be little argument about whether the time needed to author with REDEEM was worth it (estimated in this study to be between 3-5 hour to an hour). In the absence of this degree of effect, 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 students categories led to significantly better learning outcomes (though there is also no evidence that its harmed learning), 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 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 basics 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. It is true that REDEEM ITSs are not as intelligent as many ITSs and furthermore teachers have tendency to try and use REDEEM in such a way as to make it less smart (*e.g.* by attempting to prescribe a fixed prerequisite structure, using fixed not performance related categories). Potentially, REDEEM needs to monitor learner behaviour more sensitively to become more adaptive to learners' needs. We are also exploring the role of the learner in authoring environments. For example in the 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. We see a number of future directions that can profitably be addressed using a combination of pre-existing courses, ITS authoring tools, and teachers and learners knowledge.

## 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 authors and their respective schools in these studies without whose help none of this research would have been possible. We would also like to recognize the help 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 development of the REDEEM project, especially Jean Underwood and David Wood. 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*(1), 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.), *Artificial Intelligence in Education - Open Learning Environments: New Computational Technologies to Support Learning, Exploration and Collaboration* (Vol. 50, pp. 93-100).
- 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.), *Intelligent Tutoring Systems: Proceedings of the 5th International Conference ITS 2000* (pp. 182-191). Berlin: Springer-Verlag.
- Ainsworth, S. E., 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.), *Intelligent Tutoring Systems: Proceedings of the 6th International Conference ITS 2002* (pp. 873-882). Berlin: Springer-Verlag.
- Ainsworth, S.E., Williams, B.C & Wood, D.J. (2001). Using the REDEEM ITS authoring environment in naval training. In T. Okamoto, R. Hartley, Kinshuk, & J.P. Klus (Eds.). *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, pp 189-192. IEEE Computer Society, Los Alamitos, CA. ISBN 0-7695-1013-2.
- 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). Macroadapting Animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism., *Proceedings of the 5th International Conference ITS 2000* (pp. 574-583).
- 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 annual meeting of the American Educational Research Association, Chicago, IL.
- Chung, J. & Reigeluth, C.M. (1992). Instructional prescriptions for learner control. *Educational Technology* 32(10), 14-20
- du Boulay, B. (2000). Can we learn from ITSs? In G. Gauthier & C. Frasson & K. VanLehn (Eds.), *Intelligent Tutoring Systems: Proceedings of the 5th International Conference ITS 2000* (Vol. 1839, pp. 9-17). Berlin: Springer-Verlag.
- Graesser, A. C., Person, N. K., Harter, D., & Group, T. T. R. (2001). Teaching Tactics and Dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, 12, 257-279.
- Hsieh, P. Y., Halff, 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. (1995). Modelling Teaching Strategies. *Journal of Artificial intelligence in Education*, 6(2), 117-152.
- 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), 317-340.
- 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.), *Artificial Intelligence in Education - Open Learning Environments: New Computational Technologies to Support Learning, Exploration and Collaboration* (Vol. 50, pp. 191-198).
- 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.
- Shute, V. J. (1992). Aptitude-treatment interactions and cognitive skill diagnosis. In J. W. Reigan & 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.

- 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 for educational communications and technology* (pp. 957-982). New York: Simon & Schuster . Simon & Schuster.
- 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., Underwood, J. D. M., & Avis, P. (1999). Integrated Learning Systems in the Classroom. *Computers & Education*, 33(2/3), 91-108.
- Wood, D. J., & Wood, H. A. (1999). Help seeking, learning and contingent tutoring. *Computers & Education*, 33(2/3), 153-1770.