



University College London
Department of Computer Science

**EVALUATION OF COMPUTER AIDED INSTRUCTION:
ASSESSING THE VALUE AND EFFECTIVENESS
OF OPERATIONAL SYSTEMS**

Arif Mahmud Iqbal

A thesis submitted for the degree of
Master of Philosophy
in the University of London

July 2003

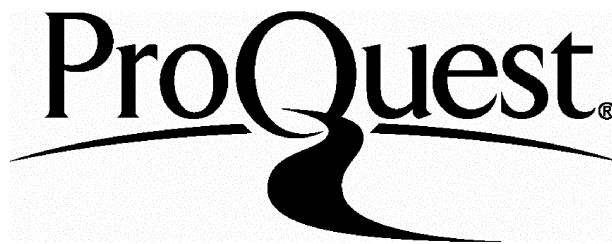
ProQuest Number: U642301

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest U642301

Published by ProQuest LLC(2015). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code.
Microform Edition © ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

This thesis investigated a number of performance measures for computer-aided instruction (CAI) systems. These "evaluation metrics" are intended to assess the worth and value of teaching systems. An operational accounting tutor (which teaches marginal costing) was used to develop the metrics and a replication study was conducted on Application Program Tutor (a tutoring system designed to teach courses). Although, CAI is a mature technology which has evolved in a variety of fields and forms since the 1950s, its potential remains untapped. Factors attributed to this include resistance from teachers, lack of student involvement in design, and insufficient imagination in curriculum design. Inadequate system standards and a deficiency of good software tools, lack of documentation, maintenance and education value have also been key limiting factors. The overall picture seems akin to a cottage industry than a co-ordinated enterprise. Evaluation is significant, to developers and users in this field, because in the short-term it improves the usability and life-span of the numerous systems that have been developed and in the long-term it focuses attention (away from the impetus to deliver) towards issues of appropriateness and quality in system design. Different traditions of evaluation are explored, including the selection criteria used in educational technology and the impact of the quality philosophy on software engineering. This research was conducted using the Before-after Two-group design on forty-two accountancy students, where a conventionally taught group was compared with the accountancy tutor group. Performance on a number of marginal (or variable) costing problems was measured before and after both groups were taught. Moreover, the experimental group was given a questionnaire to complete (which was designed to capture their assessment of the system). The results derived from the well-crafted questionnaire were indicative of the system's strengths and weaknesses and supplied useful criteria for future research.

DEDICATION

I would like to thank my parents, Muhammad Iqbal and Sharifan Iqbal
for their love and support. I would also like to dedicate this thesis to them,
especially my mother since she was my original teacher and guide (and continues to be).

Arif Iqbal,

September 1996.

ACKNOWLEDGEMENTS

Although my name is the only one on this work there are many other people who directly or indirectly contributed to the creation of this thesis, and just as it takes a team to make a dream come true, so I must acknowledge all my colleagues and seniors who made it possible.

I certainly could not have completed this thesis without the support of my supervisor, Professor Paul Samet and many thanks are due to him for taking me through both *the agony and ecstasy* of the research process. I would also like to thank the following:-

Kinshuk and Ashok Patel of the Department of Accountancy at De Montfort University for their generosity;

Paul Wernick for introducing me to the Marginal Costing system and Wanderley Lobianco for his help in proof-reading and advice;

and credit is due to John Cook (at Thames Valley University) and Mark-Elsom Cook (at the Electric Brain Company) for providing help and the loan of their system.

Many thanks for the help given by the students who participated and the permission provided by Julian Lavis of the Faculty of Business, Management and Social Studies at University of Westminster, Ian Potts of the School of Accountancy at Thames Valley University to carry out the study and loan the systems for the study and especially for Andrew Scott of the Management Centre at University College London for helping out with the study.

Jack Levy and Alan Shaw both of the Information Systems Division at University College London who were instrumental in making my disks run on the hardware used in the study and Nafisa Taylor for helping me with WordPerfect (Version 5.1).

I am also indebted to my brother for his financial support and encouragement through *the good, bad and ugly* periods of my writing up.

I am consistently reminded of that "Fear can hold you prisoner. Hope can set you free." The work carries the warning "Brains are not included - but are necessary".

There are many others, too numerous to mention, including some of my research colleagues within the Department of Computer Science most notably Kapalandu Pal, David Fulton and Angela Sasse who have bestowed their friendship and advice, I am extremely grateful to these unsung heroes. Three special individuals have had the misfortune to call me their friend during my research work. All of them shared my sorrows and happiness and two other things bound them together - a steadfastness belief in me and a true friendship as well as a first name that began with the letter "K" (i.e. Khurshid, Kinshuk and Kamran). I hope they know how much they mean to me (especially the last person), as the British poet, Alexander Pope reminded us "In every friend we lose a part of ourselves, and the best part".

This research was carried out under a Science and Engineering Research Council studentship.

Arif Iqbal,

September 1996 (revised July 2003).

CONTENTS

ABSTRACT	2
DEDICATION	3
ACKNOWLEDGEMENTS	4
CHAPTER ONE	14
1.1 The research problem	14
1.1.1 The Problem Outlined	14
1.1.2 The Gap in the Literature	14
1.1.3 Why is it important?	15
1.2 Focus of this research	15
1.2.1 Scope of this thesis	15
1.2.2 The Main Objectives	17
1.2.3 The Research Approach and its Boundaries	19
1.2.4 Significance of this Research	22
1.3 Structure of thesis	23
1.4 Author's note	26
1.5 Key Points	26
CHAPTER TWO	28
2.1 The promise of powerful educational technology	28
2.2 Computers in Education	31
2.3 A brief history of CAL	33
2.4 Current realities of education and educational software	36
2.5 Brave new world of CAI - multimedia and the Internet	43
2.6 Evaluation of CAI - The Road to Better Software	49
2.6.1 Pitfalls in Evaluative Design	51
2.6.2 When does the Evaluation end and the Teaching begin?	55

2.6.3 Instructor versus Evaluator	57
2.6.4 Establishing the goals of educational software	63
2.7 Key Points	64
CHAPTER THREE	66
3.1 What are Intelligent Tutoring Systems?	66
3.2 History of ITSs	68
3.3 Examples of Classic ITSs	75
3.3.1 SCHOLAR	75
3.3.2 GUIDON	79
3.4 Conceptual and Practical Issues in Intelligent Tutoring	82
3.4.1 Importance of ITS research	83
3.4.2 Problems faced by ITS Developers	84
3.5 Why evaluate intelligent tutoring systems?	89
3.6 How should ITSs be evaluated?	91
3.6.1 Mark and Greer's (1993) Techniques for ITS Evaluation	91
3.6.2 Legree, Gillis and Orey's (1993) Procedures for Internal and External ITS Evaluation	94
3.6.2.1 Process evaluation	95
3.6.2.2 Product evaluation	97
3.6.3 Murray's (1993) Evaluation Methods for Exploratory Research	99
3.7 Overview of evaluation methods	103
3.8 Key Points	106
CHAPTER FOUR	107
4.1 Intelligent minds or intelligent machines?	107
4.2 AI hardware, tools and techniques	108
4.3 Designing artificial intelligence	112
4.4 Evaluating intelligent systems	115
4.4.1 How to carry out a "controlled" evaluation	116

4.4.2 Informal evaluations - the alternative to "controlled" evaluation	123
4.4.3 Cognitive evaluation	127
4.4.4 Issues in evaluating	130
4.5 Landscape of evaluation	139
4.6 Key Points	140
CHAPTER FIVE	142
5.1 What is Evaluation?	142
5.2 Why Evaluate at all?	145
5.3 Formative versus Summative Evaluation	147
5.4 Models of Evaluation	148
5.5 A Brief History of Evaluation	153
5.6 The Style of Evaluation Chosen for this Study	156
5.7 Key Points	159
CHAPTER SIX	160
6.1 Introduction	160
6.2 What is Quality?	162
6.3 Quality versus Productivity	165
6.4 Total Quality Management	168
6.5 Measuring for Quality	171
6.5.1 Standards and Benchmarks	171
6.5.2 Quality Metrics	173
6.6 Key Points	177
CHAPTER SEVEN	178
7.1 Introduction	178
7.2 Determining the evaluation	178
7.3 Approaches to Scientific Research	179
7.3.1 Experiments	179
7.3.2 Basic Experimental Designs	181

7.3.2.1 Independent Measures Design	181
7.3.2.2 Repeated Measures Design	183
7.3.2.3 Matched Subjects Design	184
7.3.3 Other Experimental Designs	185
7.3.4 Correlational Studies	191
7.3.4.1 Naturalistic Observation	192
7.3.4.2 Case Study	194
7.3.4.3 Survey research	195
7.3.5 Quasi-experimental designs	196
7.4 Benefits and Drawbacks of Experiments and Correlational studies . .	197
7.5 Choice of Research Design	197
CHAPTER EIGHT	200
8.1 Introduction	200
8.2 Obtaining a fully-functional tutoring system	201
8.3 Learning about marginal (variable) costing and testing the software .	204
8.3.1 Derivation of Attitude Items	205
8.3.2 Collection of demographic and background information . . .	207
8.3.3 Piloting the questionnaire	207
8.3.4. Improvement and refinement of questionnaire appearance and layout	209
8.4 Investigating assessment tests for marginal costing	210
8.5 Contacting and arranging meetings with accountancy departments . .	212
8.6 Carrying out the study at each site	214
CHAPTER NINE	216
9.1 Introduction	216
9.2 Descriptive Statistics	217
9.2.1 Description of the Participants	217
9.2.2 Participant's assessment of the Marginal Costing System . .	219
9.2.3 Responses to the Open Questions	234
9.3 Inferential Statistics	236
9.3.1 Pearson's Correlation for inter-rater reliability	236

9.3.2 Comparing average scores of the experimental and control group	237
9.3.3 Independent (Unrelated) t-test to test the difference between groups	238
9.3.4 Cronbach's alpha - reliability of the attitudinal scales	239
9.3.5 Factor Analysis of attitudinal scales	241
9.4 APT Results	243
9.4.1 Description of APT Participants	243
9.4.2 Participant's assessment of the APT	244
9.4.3 Inferential Statistics	245
9.5 Sample Size and Statistical Power	246
9.6 Conventional Teaching of the Control Groups	250
CHAPTER TEN	252
10.1 Objectives of this Research	252
10.2 The research findings	256
10.2.1 Marginal Cost Tutor findings	256
10.2.2 Replication with Application Program Tutor	258
10.2.3 Previous Marginal Cost Tutor studies	260
10.2.4 Management Accounting Module	261
10.3 Conclusions drawn	265
10.4 Future Lessons Learnt	273

A Questionnaire	275
B Marginal Costing Tests	289
C Frequency Distribution of Questionnaire Responses	296
D Summaries of Responses to Open Questions	339
E Inter-Rater Reliability of Scores	343
F Comparison of experimental and control differences using Independent T-tests	345
G Reliability of the Attitudinal Scale using Cronbach's Alpha Coefficient of Reliability	348
H Factor Analysis of Accountancy Knowledge/Teaching, Calculator facility and Comparison with other Media	356
I Percentage of participants from the different universities	359
J Age distribution of participants (in years) and Order of birth distribution of participants	361
K Responses to "The calculator facility was easy to use"	363
L Responses to "I found it hard to use the calculator to transfer the answer to the program"	365
M Responses to "Using the calculator is frustrating - it does not do what it's supposed to do "	367
N Scattergram of pretest scores marked by two raters A and K and Scattergram of posttest scores marked by two raters A and K . . .	369
O Average group scores before and after Computer or Classroom Instruction	371
P Screen Dumps	373
Q APT Frequency Distribution of Questionnaire Responses	378
R APT Summaries of Responses to open Questions	412
S APT Inter-Rater Reliability of Scores	414
T APT Comparison of experimental and control differences using Independent T-tests	416
U APT Screen Dumps	418
REFERENCES & BIBLIOGRAPHY	424

LIST OF FIGURES

Figure 3.1: General ITS architecture (from Nwana, 1990)	68
Figure 3.2: Continuum of CAI types (adapted from Yazdani, 1986).	72
Figure 3.3: A dialogue with SCHOLAR (from Carbonell, 1970).	76
Figure 3.4: ITS Domains (from Nwana, 1990)	86
Figure 3.5: Classification chart of evaluation techniques.	104
Figure 3.6: An applied holistic model for the evaluation of ITSs (from Bento, 1990).	105
Figure 5.1: Hierarchy denoting the quality of an expert system (from O’Keefe and O’Leary, 1993).	144
Figure 5.2: Tyler’s curriculum model (from Hopkins, 1989).	149
Figure 6.1: Quality Control.	164
Figure 7.1: The Simple Experimental Design.	179
Figure 7.2: The A-B-A Within-Subjects Design.	186
Figure 7.3: The Before-after Two-group Design.	187
Figure 7.4: The Four-group Before-after Design.	189
Figure 7.5: Example of a Time Design.	190
Figure 8.1: Representation of the procedure of this study	201
Figure 8.2: Equations defining margin of safety.	205

LIST OF TABLES

Table 2.1: A history of computers in education (from O'Shea and Self, 1983). . .	36
Table 2.2: A comparative summary of educational media (adapted from Faint, 1994).	48
Table 2.3: Characteristics of Effective CAI (page 23, Hannafin and Peck, 1988). .	49
Table 2.4: Potential pitfalls in evaluative research (from Duncan, 1993 & Ransdell, 1993).	52
Table 3.1: A reasonably comprehensive list of ITSs and environments (from Nwana, 1990).	74
Table 3.2: Suitability of evaluation techniques (from Mark and Greer, 1993)	92
Table 3.3: Evaluation methods for exploratory ITS research (from Murray, 1993).	101
Table 5.1: Some models of programme evaluation (adapted from Herman et al. 1987).	155
Table 6.1: Quality assurance procedures.	164
Table 7.1: Example of subjects in an independent measures design.	181
Table 7.2: Example of subjects in a repeated measure design.	183
Table 7.3: Example of subjects in a matched subjects design.	184
Table 7.4: Comparison of experimental and correlational studies.	197
Table 9.1: What features did you find contributed to the STRENGTH of this system?	234
Table 9.2: What features did you find contributed to the WEAKNESS of this system?	235
Table 9.3: If there was ONE outstanding improvement you could recommend to this system, what would it be?	236
Table 9.4: Improvement in overall group mean scores	238
Table 9.5: Independent t-test.	239
Table 9.6: Cronbach's alpha of grouped attitude items from questionnaire.	241
Table 9.7: Improvement in overall group mean scores.	246

CHAPTER ONE

INTRODUCTION

"So we had better face the question of technology - what does it do and what should it do? Can we develop a technology which really helps us to solve our problems - a technology with a human face?

The primary task of technology, it would seem, is to lighten the burden of work man has to carry in order to stay alive and develop his potential. It is easy enough to see that technology fulfils this purpose when we watch any particular piece of machinery at work - a computer, for instance, can do in seconds what it would take clerks or even mathematicians a very long time, if they can do it at all. It is more difficult to convince oneself of the truth of this simple proposition when one looks at whole societies."

[Schumacher, 1973] (9;20;9;19;23;9;4;5)

1.1 The research problem

1.1.1 The Problem Outlined

Computer systems are increasingly being designed for the purpose of education. Regrettably however, due to the high turnover rate, an understanding of their contribution to this field has not advanced nearly as rapidly. Consequently, the need for careful and systematic scrutiny of these systems is becoming a critical factor for the evolution of this area. The overall purpose of this research is to determine a way of judging the merit or value of operational teaching systems as well as assessing their performance.

1.1.2 The Gap in the Literature

A major reason for this research is very clear, even from a cursory glance of the literature, namely that while computer systems for teaching are being developed on a regular basis for numerous domains (some of them no more than prototypes). However, little effort is being taken to guarantee a long-term strategy for their survival. This culture of technology-driven development and rapid application development is ensuring immediate

obsolescence for much of the valuable investment in teaching systems and a short-sighted approach for the whole field. This has been characterised by the lack of appropriate, or even the absence, of evaluation within this field, which in turn is hampering its progression to a mature research area and has surfaced as a considerable obstacle.

1.1.3 Why is it important?

Thus, we have a landscape of educational computing which suffers from being dominated by application rather than promotion of better practice and developing theory. It means that needs of the end-users are failing to be addressed, the expertise of users, educators and developers are not being properly utilised and not enough priority is being given to developing a learning culture which stresses refinement in design. Instead, there is a tendency to be consumed by the momentum of ever changing technologies. Consequently, this research is being carried out in a climate where there is a great need for progress in conducting evaluation studies which clearly focus on learning about the virtue and properties of current systems and their efficacy. The accumulation of such knowledge will advance the level of development in teaching systems.

1.2 Focus of this research

1.2.1 Scope of this thesis

The terms of reference for this research are defined and constrained by the following parameters:-

- a Computer Science perspective,
- an absence of a common knowledge base on evaluating fully functional teaching systems,
- a framework preventing execution of long-term or extensive set of studies.

The basis of this work originates from a Computer Science department. Consequently all the core issues raised are firmly rooted in a purely technological outlook. Our primary mission, in common with our fellow computer scientists, is to improve the current state of systems development in the promise of advancing the study and application of future computer systems, in our particular case teaching systems. This implicitly means using the scientific method and the engineering paradigm. Any contribution of knowledge in a physical science, such as computer science, relies on a variety of techniques from mathematical proof, measurement, statistical modelling to system construction in understanding and replicating physical systems. On the other hand, the role of an information systems engineer (which is comparable with other engineers who build ships, bridges and airframes) is to use their technical knowledge and skill to design and build the best system, ensuring that it is developed on time, within budget and does exactly what it was intended to do.

As a consequence, conducting research in computer science means that our principal focus is restricted to recommending changes in the development of computer aided learning systems. Thus, we are not interested in suggesting improvements to conventional teaching practice, neither are we concerned in knowing the cognitive processes that take place when using teaching systems (since we are not commissioned to produce any new cognitive model of learning). Our work solely concentrates on determining whether learning has taken place (not necessarily examining the nature or type of learning involved) and whether the chosen media was successful in carrying out this purpose (and thereby enabling us to determine in what ways it could be improved). The key point here is that all the issues we have raised are from a computer science perspective, particularly from software engineering stance.

The current state of educational computing lacks a sufficiently well developed body of knowledge (i.e. theories, models and methods) which deals specifically with the evaluation of operational teaching systems. For such an area, still in its infancy, there is a prerequisite to assemble and develop ideas and methods from other suitable branches of literature (i.e. to invoke induction). We are compelled, therefore, to turn to related fields (e.g. education, psychology and evaluation) to establish a foundation for our work. Such

a wide acquisition of knowledge can also be justified by the multidisciplinary nature of automated teaching systems, which even from its earliest origins incorporated ideas from other fields. However, in spite of the inclusion of this mosaic of literature, our sole focus is still to investigate methods that will lead to improvements in teaching systems from the computer science viewpoint.

While not a central theme of this work, the extensive examination of the multidisciplinary literature review, included a long and hard look at the human-computer interaction (HCI) and the ergonomics field (e.g. McGraw, 1992, Downton, 1993, Wilson and Corlett, 1990 and Lindgaard, 1994). This analysis of HCI models, techniques and practices failed to provide any significantly useful or unique concepts over and above those already present in the accumulated literature for this particular teaching system research. In fact, one notable reference, Howard and Murray (1987), only served to provide a confirmatory basis to the wide range of evaluation techniques available to any computer science researcher (e.g. the use of questionnaire design).

1.2.2 The Main Objectives

The purpose of this work was to investigate a number of measures to assess the *value* and *effectiveness* of *operational* computer-aided instruction systems with intended users.

To achieve this advanced level of understanding means that our major objectives are to:-

- utilise an already working system,

There are a number of good reasons for selecting a fully functioning (or operational) teaching systems for the subject of the investigation. Legree and Gillis (1991) pointed out that performance evaluations should focus on “extensive systems” since they are designed for actual requirements as well as covering a large amount of course material and as a consequence they have a substantial effect on course attainment. In addition, summative or final system evaluations are the preferred style for assessing external product attributes (see Legree, Gillis and Orey, 1993) whereas, incomplete systems are normally evaluated

formatively for internal product attributes. An unhelpful consequence of attempting to use a prototype or limited system is that the investigator can not be sure that a weakness found by using one aspect of the system was the “direct” result of the deficiency or even absence of another module. While an additional benefit of an operational system is that real users and a course are connected to its use. Indeed, Shute and Regian (1993) advocate that teaching systems should only be evaluated with the target population since they are best suited to determining the purpose of the system. So a representative system (as opposed to a working model) is the most suitable subject to use in order to demonstrate overall performance.

- find out if it is effective,
(i.e. is it at least as good as the equivalent conventional method?)

A major flaw in the literature of teaching systems to date is how so few of them have been subjected to rigorous, controlled evaluations. Shute and Regian (1993) indicates this is primarily due to the lack of appropriate control groups. Legree and Gillis (1993) contend that since traditional instruction is less expensive than computer systems, therefore if any rational decision is to be made in favour of computer teaching systems then the onus should be to demonstrate that the information technology option is at least as effective. Moreover, the comparisons between the teaching system and human tutor effectiveness is useful because it provides an estimate of how much pedagogically important aspects of human tutor interactions have been encoded into the teaching system. Murray (1993) reports that comparison studies are useful in seeking to verify program functionality, behaviour or educational effects against a well-known successful case, a new approach, or the simulated behaviour of another system. Moreover, Legree, Gillis and Orey (1993) argue that external product effectiveness is a function of a teaching system’s impact on students and therefore conclusions of overall effectiveness are supplied by that comparison.

- elicit key properties that define it as valuable system
(i.e. discovering lessons for future systems)

Winne (1993) defines an evaluation as the activity that provides key information with which to gauge the worth or value of an instructional enterprise. He argues that the growth of future work in this field relies on better evaluations and these should be judged according to how much they help the user. Underwood and Underwood (1990) suggest that the goal of software in the classroom is to assist with the intellectual development of its user. As a consequence a key aim of evaluation is to discover the major characteristics of the teaching system which facilitate this role. Winne (1993) recommends the selection of these standards of worth or value should be derived from the clients or consumers. The assembled properties benefit the refinement of design as well as the representation of the current teaching system.

- develop a procedure which could accomplish these goals easily and effectively, for evaluators.

Therefore there is no doubt that the preferred evaluation approach should include:- (a) the intended users as participants, (b) use of fully operational system, (c) the requirement to obtain learned outcomes indicating effectiveness, (d) to make comparisons with incumbent forms of instruction, (e) to acquire user's characteristics, (f) to elicit system strengths and weakness which provide clear design improvement targets. Not only that but this activity should be carried out in the most systematic and time-efficient manner as possible. The best and only reason for conducting an evaluation is the collection of valuable information which leads to immediate and much needed improvements and certainly not to become a tiresome, long-winded, pedantic exercise.

1.2.3 The Research Approach and its Boundaries

The decision to select the research design for this work was strongly influenced by the current state of the teaching systems literature (in particular the poor state of evaluation) as well as the research objectives. The largest single criticism levelled at the current state

of teaching systems is the lack of “professionalism” behind the construction of systems for teaching. Evidence for great strides in better teaching systems simply turn out to be working models (not even semi-complete systems). It can also be clearly seen in the failure to observe the normal practice of disseminating acquired wisdom to peers and an absence of a coherent strategy to employ any standard tools, techniques and methodologies. As a result, a majority of teaching systems consist of discrete, discontinuous and one-off developments, normally associated with the case study approach to construction and assessment (known in computer science as rapid application development, or RAD).

In addition, this style of research has largely contributed to the poor state of the evaluation in this field, often through idiosyncratic conclusions being drawn from untested assumptions and informal impressions and by the views of a narrow set of users. The legacy of this approach is the abundance of one-off “evaluations” using either loose, implicit and poorly-designed sets of checklists (e.g. Jones and McCormac, 1992 reported that use of user’s opinions on their own were not a satisfactory method of evaluating teaching programs, in fact prior experience proved to be better evidence) as well as the use of in-depth descriptive case studies highlighting specific situational factors (e.g. Watson, 1993 found that the positive outcomes of IT measured in longitudinal case studies were related to the teachers who set up the environment). Another nail in the coffin of this style of research is that despite it being a well-trodden path, it still suffers from a low payoff (particularly from the computer science perspective) even though the input is painstakingly laborious (e.g. Perciful, 1993 took “two years” to conduct a study on one computerised component of a nurses course just in order to address the lack of knowledge about planning, implementing and evaluating with computerised teaching programs for nurses).

The challenge now is to replace this paradigm by a more systematic and informative evaluation, which ensures large rewards in terms successive, progressive and evolutionary style of system development. Our objectives certainly shared some these key considerations, namely that the teaching system studied should be an operational one assisting us to qualify truly as a course-attached software evaluation (i.e.

representativeness) as well as providing real end-users. Accordingly the design could be comparative. The other aspect is related to the nature of evaluation which shares the attributes of the two-headed Roman god Janus, namely reflecting on what has gone on in the past and peering ahead to the future. In our case it must be used to verify what the teaching system can do and at the same time point out areas of improvement. Here we choose to measure the characteristics of worth or value from a computer system point of view. So the chosen design will reflect this balance between confirming actual performance and disclosing future potential. Once these two primary conditions have been met we can acknowledge that the research has been successfully concluded.

The actual details of the research design are discussed in chapter 7, however two main points need to be borne in mind which serve to underline the main approach chosen for this research. Firstly, in a world of ever changing technology, there is a very important requirement for educators to judge between the genuinely innovative technology and the marketing hype and a significant factor in this is decision-making is on the basis of clear effectiveness. This is a reason why comparative studies are not that uncommon in the literature. It is also a way of establishing overall performance and whether the system justifies resources to be spent for its improvement. In fact this is also the premise of benchmarking commonly used in computer science. Additionally, as we shall see, complete and working systems are conventionally evaluated by summatively evaluation i.e. the efficacy of whole system with rigorous and formal methods. While internal components and prototype teaching systems are assessed formatively, in fact Bento (1990) describes this as collecting preliminary data through informal interviews, surveys, and protocol analysis.

Secondly, most of the disappointment faced in this field has been with the lack of continuity in system development and especially the failure to learn from the mistakes of the past. It is crucial, therefore, that this trend be reversed especially since yesterday's failures can be tomorrow's overnight successes. Although it is a cliché that we do learn more from our mistakes it is more than often true. Most of the problems arise because the computer systems are so complex to develop that it can be extremely difficult to plan any systematic and thorough evaluation. This is one reason why the inclusion of a quality

management approach (from software engineering) is a vital ingredient of this work. The main purpose of quality is to make sure the system conforms to customer requirements and avoids wasted effort i.e. rework (incidentally this aspect is related to software reuse). It does this by making these requirements explicit (i.e. manifest, tangible and measurable) and then by keeping them under control. In addition, quality methodologies are often developed as products. Typical key issues include: how do existing systems meet customer needs and how to anticipate customer needs of the future and system development in the light of these needs? Since complex systems will evolve from simple systems much more rapidly if there are stable intermediate forms (than without such forms) because they can be improved and measurement of quality allows us to track that improvement, therefore it is an ideal way of tackling major improvements in teaching systems. We must emphasise that our work is beginning of a new chapter in teaching system evaluation and very far from being a developed area, but there are now a few signs that researchers in other domains are using similar methods, for instance in knowledge based systems evaluation Juristo, Mate and Tovar (1997) have suggested a quality framework.

1.2.4 Significance of this Research

The intended benefits of this research work will be to:-

- permit designers and developers to make “accurate assessments of the contribution” of their technologies on educational performance, thereby enabling the shift away from a technological emphasis towards a more requirement-driven development, through user-participative design;
- encourage more “independent evaluations” of this type for all kinds of systems, permitting educationalists to make a more informed and rational choices in selecting appropriate educational technologies;
- provide a “framework” in which questions can be invited about the design (e.g. whether there is any value added teaching) and explanations can be sought of why the system behaved so badly (e.g. because it was difficult to learn and use);

- offer “a set of tools and activities” for their own product testing (since other evaluative researchers will find it a useful introduction to a highly complex and multidisciplinary field);
- motivate other evaluators to add and to refine “the properties and the methodology” for their own systems, facilitating the development of a theory-based field dealing with evaluation of teaching systems (hence opening up the debate from a simple teaching and learning functionality).

1.3 Structure of thesis

This thesis consists of a review of several aspects from the three disciplines (i.e. computer science, education and evaluation) and particularly what they have to contribute to the debate about satisfactory evaluation approach and methods. Although this research intention is to tackle a practical problem, but in addition it needed to address the wider issues relating to the critical absence of evaluative theory as applied to this area. To some extent, the computer-aided instruction area can considered analogous to the early automobile industry, namely there are lots of different types of cars (in our case systems) being designed and built but no systematic effort takeing place improving their development. Chapter two until chapter six can be read separately as introductions to this field, nevertheless they are intended to be a cumulative set of arguments in the progress of the evaluation discussion.

In the next chapter we start by describing how computers have been promoted in the education system in spite of failing to live up to the (inflated) expectations that were built up by their early developers. We stress how these lessons should be heeded. A number of such issues are developed, including the problems of introducing computers into the classroom and making the best use of this technology, as well as whether these computers can take over some of the duties of human tutors or even if they are cost-effective. The

significant point is that whatever educational technology is used there must be careful determination of requirements and evaluation of that technology.

Chapter 3 provides a brief overview of the progression from computer aided instruction to intelligent tutoring systems. We relate the nature of intelligent tutoring systems and how to distinguish them from traditional computer aided instruction. There is a detailed examination of the problems faced by developers of intelligent tutoring systems. At this point we specify why evaluations are so important and the reasons they are rarely carried out. We report a number of methods for evaluating intelligent tutoring systems. Then we conclude by proposing a framework for classifying these methods of evaluation which overcomes some of the difficulties of choosing a suitable method of evaluation. It involves asking two questions about the evaluative study (i.e. whether it is intended to be an *internal* or *external* evaluation and whether it is an *exploratory* or *experimental* type of research) and on the basis of this a comparable technique is chosen from a chart (representing a number of techniques sorted into four main quadrants - using the two main questions as two dimensions).

We take up the idea of evaluating intelligent tutoring systems in chapter 4 and examine three different strategies how this can be carried out. The evaluation approach taken in this research reflects a mixture of these three types of evaluation. Then an explanation is given of how five broad questions help advance a high quality evaluation. Whatever approach is taken, four key guidelines are proposed:

- (1) evaluation must gauge worth or value rather than meeting targets;
- (2) a variety of data and from as many sources as possible must be collected;
- (3) there must be an element of theory or model guided evaluation and
- (4) the evaluation approach must be one that helps the user.

The discipline of evaluation is dealt with directly in chapter 5. The definition of evaluation is examined and found to be a rather confusing one. So that while the lowest level of evaluation can be considered to be a *test of correctness* (or error) and is carried out by developers and engineers, the higher level of evaluation deals with *choice or value* and

is usually conducted by managers and end users. The significance of evaluation is provided, as well as the two main forms i.e. *formative* and *summative* evaluation, and an account of how evaluation relates to education. A brief history of evaluation is outlined with a variety of models (including the style of evaluation chosen for this study).

In Chapter 6 we introduce the philosophy of quality, which in recent years has gained popularity within industry, and is now being applied to software engineering. Despite the fact that quality has a complimentary role, it is often a neglected subject when discussing evaluation. We describe why this approach is significantly better than conventional measures of productivity and why quality is not just an add-on feature of any system but is a key function throughout the entire system life cycle. The main purpose of quality is to make sure the system conforms to customer requirements and is intended to avoid wasted effort (i.e. rework), and it does this by making these requirements explicit and by keeping control of this measurement. Quality metrics is then an ideal means of tackling major improvements in computer aided instruction.

An account of how the design for the research was selected appears in chapter 7. The overall goal of this research is to develop some measures of a *fully functioning* computer aided instruction within an *operational* environment and two slightly different aims follow from this. Firstly, whether the system is at least as effective as conventional teaching, and secondly, which features (or criteria) of the system are key in assessing the quality of a teaching system. As a consequence we use two different styles of research design for this work i.e. experimental design and a structured questionnaire. Another major element, which had implications for the choice of research design, was the applied teaching environment in which the study was conducted, the difficult choices are explained.

A *personal* account is given in chapter 8 of preparations for the study and how the study was carried out. This included developing and piloting a questionnaire, creating an assessment test and arranging for a study at end-user locations. A number of practical difficulties are depicted including making a change in the structure to the before and after two-group experimental design of the research so that some of students were not deprived of the teaching. It illustrates the kind of obstacles faced by applied researchers conducting

studies in a proper teaching schedule (i.e. practicality of design). There is also an account of the hurdles faced when attempting to obtain a fully functioning tutoring system.

The results of the practical study are presented in chapter 9. There is profile of the participants describing their demographics and their current and past experience with computers. This is followed by a detailed analysis of both groups performance on the tests and the experimental group's assessment of the Marginal Costing Tutor.

Chapter 10 outlines a discussion and an interpretation of the results. It also reviews the contribution of this work and the practical lessons to be learnt from this type of investigation. We have provided a number of conclusions and added weight for the need to carry out further work. Evidence of other similar work is highlighted to emphasise the significance of such evaluations to this area. We end by suggesting a number of research awareness that remain unexplored.

1.4 Author's note

Please note the term *we* (used throughout this thesis) is simply a writing convention.

The quotation convention we have followed in this thesis is that long quotations from the work of others are enclosed in double quotation marks, while small phrases (usually less than three words) written for emphasis are designated by italics or double quotation marks. The reason for using direct quotations is that the authors' words and meaning are preserved from the original source (although it is acknowledged that when a reference is out of context readers may take a different meaning from the original). The remainder is to be attributed as the work of the author.

1.5 Key Points

- Computerised teaching suffers from a rapid turnover in system production, technical know-how and resource management.

- It has become necessary to establish a new culture based upon successive system development which is user-focused, easily auditable, and ensures large communication of knowledge.
- The basis of this computer science work is to pioneer an alternative evolutionary approach using the best of current evaluation practices and the application of some ideas in quality management in order that future evaluators are able to build upon this approach to refine teaching system design.
- To demonstrate this approach we intend to assess not only the value but also the effectiveness of an operational computer aided instruction system on real end-users using an appropriate evaluative methodology.

CHAPTER TWO

EDUCATION AND CAI

"In the 1960s there was much talk of a generation gap; but that was understood to be a moral and political discrepancy. In the Information Age, the gap is purely technological, a matter of programming talents and keyboard virtuosity. 'Kids and computers click', possibly in a way that leaves their parents no option but to stand aside and watch with amazement - but only after they have gone shopping and bought the equipment. Undeniably, some kids click with computers. The emphasis, however, belongs on *some* - as in the phrase, *some* kids click with violins, or *some* kids click with paintbrushes. But there are no millions being spent to bring violins or paintbrushes into the schools. Initially, there was a simple justification for favouring computers over violins in the budgetary priorities of the schools. It was embodied in the catchphrase *computer literacy* - a seemingly undeniable necessity in the Information Age. Lacking that skill, children would grow up to be unemployable." [Roszak, 1986] (12;25;1;3;3;5;16;20)

2.1 The promise of powerful educational technology

Educational technology is commonly understood to mean the application of technology (usually hardware) to education and includes everything from blackboards, wallcharts, overhead projectors, film projectors to radio, television and computers. Although the roots of educational technology can be traced as far back as the ancient Greeks the term first came into general use in the late 1960s following the heyday of "programmed instruction". Programmed instruction originated from the principles of learning developed by behavioural psychologists, most notably B.F. Skinner, in laboratories and involved teaching material in a controlled, programmed way. The ideas of programmed instruction directly led to interactive text delivered in the form of the "teaching machine" (a device which actually took over enough of the teacher's essential functions to be able to teach automatically). Skinner's considerable influence on early developments in this field came primarily from his analysis of the learning process, which lent itself relatively easily to mechanization. The attraction of programmed instruction with educators, however,

stemmed from its ability to provide an individualised self-instruction teaching package, which was also cost-effective.

However, while programmed instruction enjoyed considerable popularity in the late 1950s, the revolution forecast by Skinner and others did not take place. Instead, the drive to develop programmed self-instructional texts continued in many other forms, especially because the principles were so applicable to the newest technology i.e. the computer. In fact, ever since computers came into common usage they seemed ideally suited to education. There is a number of reasons why this should be the case for instance, their capacity to store and retrieve huge amounts data, their fast computing power and infinite patience. Therefore it is not surprising that Guthrie and McPherson (1992) reported that from the earliest "development of computer assisted instruction (CAI) on large mainframe computers in the early 1960's to its current use on microcomputers, CAI has been heralded as one of the greatest innovations in education". Another set of educational researchers, Underwood and Underwood (1990) have described some of the remarkable claims that were made, "Patrick Suppes predicted that developments in educational technology, and specifically in computer usage, would change the face of education in a very short space of time. This prophecy was made in 1966, and was based on his perceptions of the unique capabilities of the computer. He saw it as a tool which can be used interactively, presenting materials in novel ways not easily available through other media, and with the flexibility to adapt to different learning and teaching styles. The second prophet, Seymour Papert, also expresses ambitious aims for classroom computers, suggesting that we can abandon the worksheet curriculum and confidently allow children's minds to develop through the exploration of computer-simulated 'microworlds'". (5;4;20;8;1;20;14;15) Unfortunately computers have generally not lived up to these early expectations (see Rich, 1991) and in this chapter we shall explore some of the reasons for this disillusionment.

One of the reasons for this misplaced optimism was that computers simply came to be used to reproduce books in electronic form or to revive programmed instruction. The original goal of this new educational technology was to construct instructional programs that incorporated well-prepared course material in lessons that were optimized for each

student. Early programs, however, became electronic "page-turners" that printed prepared text and executed simple, rote drills; and practice monitors, which printed problems and responded to the student's solutions using pre-stored answers and remedial comments. Another reason for the disenchantment is given by Hawkrige (1991), who argues that many of the problems of this area, including the lack of any real success, are linked to its theoretical foundations which have to be challenged. Hawkrige (1992) strongly contends that the petrification of educational technology was because its theory had not moved on from the behavioural science tradition of programmed instruction. One solution provided by him is to borrow ideas from cognitive science (in the same way, as we shall see, that intelligent tutoring systems have done).

On the other hand, Atkinson (1983) notes with some surprise that, despite the fact the education industry is renowned for so many curriculum changes, there are few signs of technical innovation. He questions whether this is solely down to higher costs (in comparison to the benefits) of introducing technology or if there are other constraints on the individuals or institutions involved and goes on to say that "The use of educational technology promises higher output and/or lower costs. Greater learning may occur because different, more suitable methods can be used. The methods will vary with the medium, but several of the media such as computer-based learning, allow individualised instruction where students can learn at their own pace. Moreover, if alternative learning methods are available, students can choose that which they find most appropriate. Some of the media may be particularly appropriate to some types of learning". Another benefit of innovation is that, since the material may be better prepared than an individual teacher could manage, it could be carried to a wider audience.

Atkinson's main argument, however, relates to the broader issue of cost in relation to context of use. In schools, educational technology is often used as an extra help for teachers and they are expected to maintain control of this technology and as a result it tends to add to costs. On the other hand, in higher education, it may be used as a substitute for people, therefore reducing costs. Atkinson cites the real obstacle to technical innovation as being related to initial costs and scale, so that "in general the pattern of costs will vary considerably from those found in 'conventional' teaching. A fundamental

difference is that a media based approach demands a far greater investment of resources before any output is obtained. This is particularly true for computer and distance learning, where the fixed costs are usually very large. This means that the scale of the operation is important. If the fixed costs are large, the approach will only be economical if these can be spread over many students". He concludes that high fixed costs, such as teachers needing to spend many hours producing a tape-slide programme, is one reason why the pace of educational innovation has proceeded so slowly.

2.2 Computers in Education

Before continuing we must tackle one of the contentious issue which afflicts the literature in this field, that is the disagreement between the three main terms which describe the application of computers in education. These three terms are known as Computer-Assisted Learning, Computer-Assisted Instruction and Computer-Managed Instruction and are abbreviated as CAL, CAI and CMI respectively. As Romiszowski (1988) points out, "every author seems to have his own definitions and classifications for the same technical terms. One marked difference is in the use of the terms 'Instruction' and 'Learning' on either side of the Atlantic. In the USA particularly, the term 'Instruction' tends to be used in a more global sense, for any type of teacher/learner interchange, whereas in Britain the term CAI has of late become restricted to 'programmed instruction' types of exercises, the preferred generic term being 'computer-assisted learning' (CAL). It seems that Canada, caught between the two influences, cannot make up its mind,". He goes on to suggest that even British authors disagree amongst themselves, although most recognize that CAI is a subset of CAL. On the other hand, CMI is sometimes considered as a variety of CAL and at other times as a separate category, not part of CAL in the strict sense. To further confuse the picture, there are two versions of CAL, occasionally seen as Computer-Assisted Learning or alternatively as Computer-Aided Learning (the same applies for CAI). For the purpose of this thesis these two versions mean the same thing. The reader should also be aware that the use of computers in education has spawned a variety of other terms such as educational software, instructional software, instructional technology and learning technology as well as computer-based instruction/learning (CBI/L). As for

the three main terms we give the following definitions, based on those provided by Romiszowski (1988):-

Computer-Assisted Learning (CAL) - is a global term indicating the variety of ways that computers are being used in education and training; it is taken to include simulations, games, database search/inquiry methods and programming of computers.

Computer-Assisted Instruction (CAI) - is where the computer explicitly delivers a set of instructional programs and is associated with drill-and-practice, dialogue and tutorial forms of computer use.

Computer-Managed Instruction (CMI) - is an umbrella term which includes all the routine data processing tasks that an instructor needs to assess students or to revise materials.

Although it is hard to generalise, the use of the computer as a tutor or teacher can be classified under term CAI, whereas CAL describes the use of the computer as a tool. Computers, however, have come to be used in a variety of ways in education, each of which is related to a distinctive version of computer-assisted learning:-

Simulation - where the properties of a dynamic system or situation are reproduced, because it may be expensive or dangerous to construct for real;

Drill and Practice - the presentation of questions with immediate feedback, in order to reinforce simple skills or to practice already learnt skills;

Problem-solving - where the learner is prompted to resolve a dilemma or problem by systematically acquiring or applying lesson knowledge, and is considered a useful way to impart some elementary skills;

Dialogue - is an approach which enables a two-way conversation between the user and the computer (particularly where both are allowed to ask or respond to questions), the intention is to mimic human learning/teaching interaction;

Tutorial - is where the learner is provided with explicit instruction so that information is taught, appropriate problems are presented, and informed guidance and feedback is provided during the lesson; it is a relatively more sophisticated form of instruction (since there are remediation and strategies for making the learning more meaningful to the learner).

2.3 A brief history of CAL

Computer assisted learning began in the 1950s with *linear programs*. They are mainly associated with the psychologist B.F. Skinner and are based on the principle of operant conditioning. They consisted of computer programs which presented students with a series of *frames* (or units of information). Each frame would consist of a question and a space to make a response. The students would then be immediately informed if they were correct, before proceeding to the next frame. Nwana (1990) describes why this instruction was so basic, "The major limitations of linear programs became glaringly apparent: they did not provide individualization, which meant that all students, irrespective of their abilities, background, or previous knowledge of the domain, received exactly the same material in exactly the same sequence; neither did they provide feedback, as the students' responses were ignored". The last comment means that, apart from being told the answer was right or wrong, there was no additional corrective information. Another flaw of these linear programs was that they were only suitable for an associative level of instruction.

In the 1960s Norman G. Crowder developed a new approach called *branching programs*, which improved upon some of the limitations of linear programs. Although they still used a fixed number of frames, they did contain multiple-choice responses, which enabled the student to select an alternative so that they were "branched" to a specific next step based on that response. These programs presented a frame to the student, which was then checked, if the student's response was correct then the next frame would be presented; however if the student was incorrect then further information would be presented exploring the error and possibly repeating an earlier frame. The idea was to allow bright students to move quickly through the program and the less able students to be given

remedial material when they answered incorrectly. Moreover, unlike linear programs, Crowder presented larger amounts of information to the student (up to a full page) before requiring a response. Although Nwana (1990) does point out "the teaching material became too large to be manageable through straightforward programming and so a special breed of programming languages, called 'author languages', were developed for creating CAI material". Research indicated that these programs were effective in teaching many different types of objectives and a wider range of students. Their shortcoming was that the branching decisions were still specified by the author.

Generative systems appeared in the late 1960s and early 1970s and represented a new level of sophistication in the design of CAI systems. The gains in computer memory and speed meant that teaching material could itself be generated by the computer. Nwana (1990) relates that, "A generative system has the ability both to generate and solve meaningful problems. In some domains like arithmetic, researchers realized they could do away with all the pre-stored teaching material, problems, solutions and associated diagnostics, and actually generate them. The potential advantages, if exploited, were enormous. They included drastically reduced memory usage and the generation and provision of as many problems (to some desired level of difficulty) as the student needed". Generative systems did improve individualization and feedback and are often considered as the first *adaptive* teaching programs, since the selection of problems varied as a consequence of the student instead of the author. They did, however, suffer from a lack of explicit representation of student knowledge and therefore were unable to question the student. In addition, generative systems did not have the human-like knowledge of the domain they were teaching and therefore were unable to answer serious questions by the student as to why and how the task should be performed.

CAL systems have improved in technical complexity over the last four decades from their humble beginnings of replacing the programmed learning machines. However, in overall terms they have remained very basic. While students could work through the material at their own pace and correct responses were immediately rewarded, students' level of competence could not be assessed. It would be difficult, for instance, to know if they were simply responding in a trial and error manner. As Yazdani (1986) states, "Computer

assisted instruction (CAI) is a mature technology. The use of 'author languages' makes construction of such systems reasonably straightforward. The main problem with CAI systems is the shallow representation of knowledge of the domain in which they teach, and the fact that they are suited to teaching specific expertise and not abstract problem-solving activity". Even in generative systems, where the teaching material was produced by the computer itself, there was a limited improvement in the variation of difficulty of the task. Generative systems were the precursor to a new type of CAI called intelligent tutoring systems, which will be explored in the next chapter. Lo (1991) has traced this evolution of CAL, from the early days of linear programs to the use of intelligent tutoring systems in the 1990s. During this period, he indicates there has been a shift in the locus of control from tightly defined program control to increasing student freedom. He observed that such a shift could be seen to mirror a change from a behaviourist-oriented approach to a more cognitive, student-centred perspective to learning. While linear programs were clearly identified with behaviourist theories of learning, generative systems were a movement in the direction of the cognitive approach, i.e. meaningful instruction. We have only managed to give a brief account of the history of CAL, for a comprehensive picture one should refer to O'Shea and Self (1983). Instead, in order to illustrate the haphazard way in which ideas in CAL have developed, we have provided a list of the various CAL programs from their textbook.

Approach	Distinguishing characteristics	Illustration
Linear programs	Derivation from behaviourism; systematic presentation; reinforcement and self-pacing.	Last (1979)
Branching programs	Corrective feedback; adaptive to student response; tutorial dialogues; use of author languages.	Ayscough (1977)
assisted learning	Drill-and-practice; use of task difficulty measures; answering student questions.	Palmer and Oldehoeft (1975)
Mathematical models of learning	Use of statistical learning of learning theories of limited applicability; response-sensitivity.	Laubsh and Chiang (1974)
TICCIT	Team production of courseware; "mainline" lessons; use of television and minicomputers; learner control.	Mitre Corporation (1976)
PLATO	Multi-terminal interactive system; visual displays; "open shop" approach; concern over cost.	Bitzer (1976)
Simulation	Computer as laboratory; interactive graphics; typically small programs.	McKenzie (1977)
Games	Intrinsically motivating; audio-visual effects; often lacking educational aims.	Malone (1980)
Problem-solving	Computer as milieu; programming by children; derivation from Piaget's theory and artificial intelligence.	Papert (1973)
Emancipatory modes	Computer as labour-saving device; task-oriented; use of microcomputers and public information systems.	Lewis and Tagg (1981)
Dialogue systems	Tutorial strategies; use of natural language; mixed initiative; use of complex knowledge representations.	Carbonell (1970)

Table 2.1: A history of computers in education (from O'Shea and Self, 1983).

2.4 Current realities of education and educational software

Educational institutions in Britain and other countries, such as the United States of America, are facing huge challenges and some are referring to this state of affairs as a crisis. Symptoms of the British education crisis appear as: a teaching profession without

much recognition (often reflected in monetary rewards); an education system which is seen as to blame for an endless list of problems e.g. widespread illiteracy, school bullies and crime. Indeed a recent report (*Every Child in Britain*) highlighted the current debate about poor standards in simple mathematical and language skills among school children. More underlying reasons are giving rise to this crisis, including tighter controls on education funding, a reduction in the younger population, gaps in technical and vocational teaching and a fragmentation of society. The new age of information technology is imposing its own pressures, most notably through an increasing demand for a better qualified workforce. Consequently it is no surprise that solutions to this crisis are also being sought from this very same technology.

The most significant challenge to education comes from the Information Age. In the forward to the second edition of Papert's book *Mindstorms* John Sculley, ex-CEO of Apple Computer, highlights what this means for the USA,

"During the Industrial Age and for most of this century, America stood alone at the top of an economic pyramid, taking resources out of the ground - oil, wheat, and coal - adding its manufacturing know-how to those resources, and selling those goods to the rest of the world. We are no longer in the Industrial Age. We are in an Information Economy where strategic advantage is determined by ideas and information and by the skills of a nation's work force.

Virtually overnight, America has gone from being resource rich to being resource poor. As a direct consequence, America is perched on the edge of an economic cliff, and unless we make a concerted effort to bring the educational system into sync with the rest of the global economy, we are in danger of supplying the rest of the world with low-wage work and losing out on the high-skill, high-wage economy that the rest of the world has moved toward". (14;5;2;21;20;20;8;5)

His comments are typical and imply that education has to undergo a major reconstruction or even a complete transformation to meet this challenge.

If we look back to the origins of the present school system this is nothing new, educational institutions have always developed to meet the needs of the country's social and economic systems. Knirk and Gustafson (1986) describe how the Industrial Revolution introduced the practice of mass education, "Factory schools were established by industries in the United States in the 1800s. They arose because the apprenticeship system did not provide the necessary number of trained people for the factories. Also, because of increasing job specialization workers did not need to be highly skilled at many tasks; they usually had to know how to do only a single operation. Because the workers needed little general education, the training program could be tailored to the specific tasks to be accomplished. During that period, industrial progress was not so rapid that it was necessary to teach principles to permit generalizable understanding and the ability to work in varying or evolving situations." A similar series of events took place in Britain. In fact, some of the earliest schools were established in the Industrial Revolution to educate the children of the workforce, which had moved from the small farming and craft communities of the countryside into the huge factories of the rapidly growing towns and cities.

Currently, the education systems of the industrialised countries are attempting to solve the crisis by increasing their transparency to audit, i.e. by making decisions clearer and accessible to the public. For instance, in England and Wales this has meant introducing a National Curriculum, national school tests, devising a mechanism for parent involvement, and education inspection through OFSTED (who monitor and improve the standards of achievement and quality of education through regular independent inspection, public reporting and informed advice). Other proposed solutions have attempted to change the education system itself. An important factor here has been the trend away from the entrenched practice of mass education and towards an open learning style of education. Open learning is often misused as a buzz-word, but essentially refers to a mixture of old and new technologies designed to make learning more accessible, by allowing the learner a degree of flexibility as to the time, place and the pace at which they cover the chosen programme of study. It also implies the education or training must be learner-centered, individualised and flexible. It is a philosophy diametrically opposed to the convenience of the teacher (i.e. providing classes at a fixed manner, time and location). It also reflects

the fact that education is no longer constrained to the young and that adults are required to possess a wider education base and regularly update their skills in the pursuit of life-long learning.

A related philosophy of education is one in which the learning process is conducted from the student point of view. The modern proponent of this type of learning, known as the *child-centered approach*, is Piaget and involves a more individual and variable interaction between the learner and their environment. The Piagetian heritage acknowledges that at the heart of any good teaching and learning experience is a critical relationship i.e. where teachers and learners alike seek to question each other's ideas, to reinterpret them, to adapt them, and even to reject them but not to discount them. The most important lesson of Piaget's view of learning is that it is an *active process*; the teacher is not a mere explainer or imparter of information instead their main task is to foster conditions under which the student can think freely. It has been argued that the development of information technology will make these styles of learning easier to develop and implement. For instance the Paul Hamlyn Report on Education declared that "IT is playing an increasing role in the delivery of further, higher and adult education and training. It is making learning available to much larger numbers by giving students more choice over the time, location and pace of learning, and by securing a greater commitment to learning". Furthermore the report described a survey which highlighted the need to monitor the quality of IT applications as well as the scale of provision. Another variety of this kind of learning which is being supported by new technologies is Resource-based learning (see Taylor and Laurillard, 1995), in which the capabilities of a library are provided for a self-directed learner.

A question which naturally arises from this is whether a form of technology implementation is more likely to solve this crisis. We discussed this in a previous report Iqbal (1990), with special reference to the American education crisis. For others, however, information technology is capable of generating more radical remedies to this crisis. Most notably Seymour Papert, who has argued that the computer can act as a catalyst to fundamentally transform the way people think, work, learn, and communicate. Indeed Sculley tells us that "Seymour was among the first to see that massive change was needed

in the education system, particularly math and science education, and to recognize the role that technology could play in learning. Perhaps more significant, though, was how he was one of the first to recognize that technology in the classroom was not a silver bullet that would solve all of education's ills." It is this last point which will be discussed in this section. While CAL is not a panacea for all the ills of education (indeed it can provide problems of its own), can it provide an effective form of education? This is a crucial question and needs to be asked more frequently.

In the same way that other machines, for instance steam engines, have replaced muscle power so the arrival of the computer persuaded us into thinking that it will assist or even replace our brain power. At a time when there is a shortage of people to do the actual teaching and teaching costs by other means are rising, schools and other educational institutions have become keen to utilise this technology. It is important to say here, however, that while it is tempting to make general comparisons between human and computerised teaching systems it is also clear humans and machines have their separate advantages and limitations. Therefore proper comparisons should be limited to particular situations. Despite this obvious disparity, many computer visionaries continue to advocate that all one has to do is to throw technology at the problem. Consequently there are huge amounts of money being spent on computers in education. In fact in the USA there is a personal computer for every seven pupils, whereas in the UK that figure rises to fifteen¹. Snyder and Palmer (1986) even described a number of fashionable notions on how computers are being propelled into schools:

Comfy Is As Comfy Does - is to create an ideal children's bedroom in school in order to make the computer more accessible (i.e. keep it away from the chaotic classroom).

Face the Music - is to put the computer in the mathematics lab where it belongs (i.e. maths wizards who want to use it and they are the ones who know how to use it, so don't fight it).

¹Source: Communicate, March 1996, page 23

Close-door Policies - which make the mistake of putting the computer in the mathematics lab (i.e. preventing the rest of the kids to get so much as a foot in the door).

A Separate Peace - is not to put the computer in a corner of the classroom, because one kid will dominate the machine and upset everyone else. One might think his enthusiasm would spread to other kids, and it does occasionally, but they will have to wait their turn.

Networking for Kids - is to try to get as many terminals as possible, so each child can have a turn. You can give the whole lot of them their drill-and-practice, and march them out again. It is much easier to administer this way.

Wholistic² Computing - whatever you do, don't buy into a rigid networking system. These systems demand a huge commitment of time and attention simply to be kept in working order. A loose, almost random arrangement is preferable. It is best if computers are available to any and all children, regardless of their perceived interests or learning styles.

They assert that "Contradictory though these positions might be, they have in common a certain cart-before-the-horse quality, a hardware-drivenness that tries to fit the situation - the classroom, the kids, the lessons - to the technology". As we have seen, it is exactly this blind optimism which has led to disillusionment among educators.

In fact, teachers should be well-versed in this overoptimism and disappointment by now. As Kerr (1989) points out "For those interested in instructional materials, perhaps no feature of twentieth-century educational practice has been more puzzling than the gap between what technology has promised and what it has delivered. That the techniques for encoding, editing, moving, and displaying information have multiplied and improved over the last hundred years there can be no doubt. And, as each approach (photography, film, television, sound recording, computers, and then computer control of those earlier systems) was introduced, educators expressed their conviction that this new way of capturing and communicating information would lead to significant improvement in what teachers could

² also known as = holistic

do for learners". (2;18;1;22;5;4;5;19) As we now know each of these beliefs proved to be unrealistic aspirations. However, Kerr provides a more useful appraisal of electronic and traditional texts, particularly their role in the classroom. Moreover, O'Shea and Self (1983) were among the first of a number of researchers to recognise that problems can arise when making effective use of the computer in the classroom. They understood that it was not just related to the technology (which in fact that was the least complicated aspect) but it came from the social organisation surrounding it. They trace the problems as arising from four main sources: the reaction of the people involved; the poorly-designed materials; technical matters (such as cost and reliability) and evaluation (i.e. actually deciding whether what has been done was worth doing). More recently, Reynolds (1990) has examined why educational systems have failed to make the transition from research laboratories to the classroom and more importantly recommended approaches which should be taken to overcome this. His research attempts to deal with some of these *poorly-designed* and the unpardonably difficult-to-use software available that are typically to educational institutions.

We know that producing good educational software is not an easy task, indeed Marshall (1988) relates that "Given the range of skills to be called upon, the subsequent creation of appropriate CAL software may be better done by a team comprising subject-based specialist, programmer, graphics designer and educational methodologist. It must be acknowledged, however, that it is difficult to get such teams together and to enable them to produce software of significance. Woodhouse and McDougall believe that such teams should be attached to curriculum development groups where longer-term commitments may be made. This involves organisation and a proper financial base". As a consequence, computers should not be viewed by schools as a short cut method of teaching, like other forms of instruction the development of high quality software takes huge amounts of time, money and effort. We end this section by reflecting on Marshall's (1988) advice on employing computer assisted learning, "It is but one of many possible presentational media available to the teacher. Thus, CAL should only be used where it is educationally justifiable and not merely for its own sake. If a rational decision is to be made in favour of CAL, this necessarily involves an appraisal of the entire curriculum and existing teaching strategies". We hope that such pragmatic opinions prevail with the computer

enthusiasts since it is doubtful that ineffective instructional technology will provide any solutions to the education crisis and may even exacerbate matters.

2.5 Brave new world of CAI - multimedia and the Internet

There has been a general trend to make the newest technology smaller (i.e. portable) and to design it for home use, examples include mobile phones, electronic organisers and games modules. The most obvious case has been the invasion of the personal computer (PC) into the home. A recent Market Report³ by Olivetti declared that nearly a third (31.8 per cent) of British households have a personal computer. This is the highest level of computer ownership in the world, compared with 21 per cent in France, 20 percent in Germany, 18 per cent in Ireland and only 15 per cent in the USA. Not only that, users reported spending an average of 10 hours a week at their computer screen and this is the first time that computer usage has dented the amount of TV viewing. Young people are starting to turn off their televisions, spending up to 20 hours a week using computers for education and entertainment. The impact of these changes is far too early to tell; however, Peter Cochrane the head of advanced applications and technology at British Telecom⁴, describes how advances in educational technology have influenced his schooling and what is likely to be waiting for us around the corner,

"When I was a young boy, there were three principal books in my home: The Bible, an atlas, and a dictionary. So, for me, school (in the '50s) was a treasure-house of information, intellectual stimulation, and learning. There were books, smart people, and resources. School became the oasis, home the desert.

Today, many children experience the opposite. Schools are short of books, computers, and information, and larger class numbers severely limit access to terminals. So children more and more are discovering a computing paradise in the home - often well stocked with CDs, network connections, and challenging interactive software". (5;18;22;5;20;8;5;6)

³Source: The Guardian, Thursday 2nd May 1996

⁴Source: Wired, October 1995

These changes appear to be revolutionary (since they indicate that education is moving from schools into the home), but before we indulge in the same fantasies presented by the computer visionaries let us step back and examine the shape of instructional technology in the future.

Two new technologies are promising to influence education - multimedia and the Internet. The first thing to clarify is what is meant by multimedia. Multimedia is a computer - or a television with processing power built in - that is capable of simultaneously sending and receiving images, voice, data and moving pictures. It should also be possible to interact with the information being displayed. The key factor which has made multimedia possible is the CD-ROM (i.e. compact disc-read-only memory). This is a laser disc which can hold vast amounts of information (typically 250,000 pages) and is designed for optical playback. The important part is that software written specifically for the CD-ROM is able to serve as the interface between a CD-ROM player and a computer to facilitate rapid access to information stored on the disc. Preston (1992) describes the benefits of this technology for education, "The combination of high quality audio with video, graphics, text and data, all interleaved together and offering fast information retrieval under user control, has created a new and fascinating product, called *CD-Interactive* or *CD-I*." On its own this does not seem greatly remarkable but, an additional attraction has been the development of hypertext. Hypertext is a software concept which packages knowledge in non-linear ways, allowing the user to explore it by self-determined linkages. It means that the knowledge is not constrained to the linear-document form. Some have regarded it as a tool which allows information to be retrieved from the computer in many of the same ways that is accomplished by human memory, for instance it enables ideas to be associated quite easily and related material to be browsed.

The term hypermedia was coined to describe hypertext applications which include graphics, video or audio. Hooper (1990) describes how this hypermedia was eventually realised in the form of HyperCard and credits it as a unique technology, "To my mind the core argument was that HyperCard is a tool that is absolutely fantastic for education, be it the education of small children or adults, including adults who work in corporations. HyperCard did not fit neatly into the prevalent application categories: Strictly speaking,

it wasn't a database or a spreadsheet or a word processor or a graphics program or a desktop-publishing application - yet it didn't seem all that strange to me." The significance of this application is yet to be fulfilled, but as an educational tool Semper (1990) contends HyperCard stands out in three important ways. Firstly, it provides the possibility of a truly multimodal presentation (i.e. the control and presentation of sounds, images, text, and graphics in a dynamic or static manner under user control). Secondly, HyperCard is the ultimate do-it-yourself medium (i.e. it allows you to design screens, deal with text and images, and create your own links - all with tools built into the program). Thirdly, the structure of HyperCard mirrors some of the successful strategies for learning, and it therefore becomes a useful tool for thinking and developing reasoning. A lot of these powerful features are already giving credence to the opinion that we will see a *multimedia revolution*. Indeed the rhetoric is quite reminiscent of the emergence of computers, for instance Baker and Tucker (1990) declare, "The converging technologies of desktop publishing, hypermedia and the compact disc have given us a multi-faceted phenomenon called multimedia which seems set to supersede computer-based training (CBT) and interactive video (IV) in education and training. The term interactive learning has been coined to describe the process of learner-based learning using a multimedia approach". There is concern, however, that large publishing corporations are now getting involved in interactive books using CD-ROM and since they do not know enough about software it is unlikely to be truly revolutionary as we might imagine. In fact they maybe acquiring some of the lessons that educational software developers have learned in the last three decades.

The other technology which has the potential to redefine our education is the Internet. The Internet is a connected web of computer networks, each of which may have hundreds or thousands of computers participating. Its origins are in the USA but its use is becoming worldwide. Previously mainframe computers have often been connected to each other so that data was able to be transmitted between organisations or even countries. However the current enthusiasm for the Internet derives from the idea of sharing resources across long distances and has powerful benefits for a greater number of people. The impetus for the widespread use of the Internet is coming from the convergence of the twin technologies of telecommunications and computers and companies within each field are

eagerly promoting it. While there is a considerable number of applications available on the Internet from electronic and voice mail to desktop videoconferencing, there still needs to be a co-ordinated development of education applications and the introduction of a variety of network services for schools, colleges and universities. The students who will benefit most will be those who are too remote from resources, such as libraries and teachers (e.g. Open University students). Educational institutions may also benefit by facilitating access to materials at the times of the students choosing. The future cost of telecommunications is set to drop dramatically, optic fibre will be available for greater capacity and the downward price pressure on computing power is set to continue. Consequently, we have a situation where affordable and beneficial services are a realistic possibility. Again this depends on whether companies involved are likely to provide sufficiently high levels of service. In addition, governments will be required not only to set up access for schools, provide adequate financial resources but also develop policies which will ensure high standards and educational innovation are maintained.

Both multimedia and the Internet will be large influences on how we will learn, but the crucial aspect must be to make effective use of each technology. For that we need to understand its properties and to scrutinise its consequence. To illustrate this point we have provided a comparative summary of various educational media using information from Faint (1994) in Table 2.2. A key aspect to all new technologies is the question of access and control, i.e. who has a part to play in its use and development. Certainly there are gaps opening up which are traditionally present in society, most notably women and girls are less likely to wield powerful influence over such technologies and people in a lower class will be least inclined to have access to them. Younger people, however, are more likely to be comfortable with the technologies. The reason for emphasising this point is to demonstrate that the people who this technology is aimed at and who are likely to adapt it for their purpose will be a privileged minority. Another essential aspect is whether learners will be alarmed by or even reject the introduction of this technology. Sculley answers this question, "An important finding of our Apple Classroom of Tomorrow (ACOT) research, and a point that Seymour [Papert] makes, is that the introduction of computer technology into the classroom did not have the expected - alienating - effect that many people worried about back in the 1970s. On the contrary, what we have found is

that in classrooms saturated with technology there is actually *more* socialization and that the technology often contributes to greater interaction among students and among students and instructors". Indeed, if the use of computers and telecommunications is properly implemented it can provide the kind of education described as computer supported collaborative learning by Kaye (1995). What is crucial in this case is whether our educational institutions are prepared to make effective use of such innovations.

Whether multimedia or the Internet are successful is not altogether important for education. New technologies will appear, evolve and even disappear. What is significant is that any new educational technology must be developed and utilised in an effective manner. For instance the demand to employ technology within the home is already being impeded. A recent report⁵ noted that a lack of simple, user-friendly software packages was seriously constraining the use of home and portable computers to access business applications and information systems. Therefore the priority for educational researchers must be to discover the properties of this technology and to shape these benefits for the students, hence the need for evaluating technologies and determining educational requirements. This process is difficult and never-ending, but equally vital. Castellan (1993) is confident about the merit of evaluation, when he describes that "information technology embraces a wide range of media, including software, CD-ROM, interactive video, and sound. As I survey the new information technologies that are available to educators, I am convinced that effective evaluation is essential, and some of the lessons that we have learned about the effective use of instructional software also apply to all aspects of information technology in education". How such evaluation should be carried out is the subject of the next section.

⁵Source: Schema Report, in Voice, April 1996

Type of Media	Strengths	Weaknesses
Audio	Widespread and familiar. Equipment cheap and robust. Audiotapes are quick to distribute. Personal and intimate. Large degree of learner control.	Access to a player is necessary. Complex branching is difficult. Information is intangible. Difficult to absorb complex information. Difficult to find relevant part of tape.
Video	Flexible and easy to use. Cheap for large groups/longer time. Good at stimulating debate. Cheap to copy and distribute. Low technical skill to set up and use.	Non-interactive and passive style. High TV quality led to high expectations. High costs for top quality production. Programme can become dated. Unsuitable for complex data or skill.
CBT	Effective for sequential info. Convenient for testing knowledge. Most computers have ability for CBT. Powerful and intuitive authoring packages. Packages are relatively cheap and easy. Effective provided not complex animation. Consistent training and learning own pace. Popular due to stimulating and interactive courseware.	Text-based can be dry and machine-like. Easy to create labyrinthine monster. Hard to run on different computer. Have to decide hardware platform.
Interactive Video	Wide generic training titles available. Full-screen, motion video. Hardware platform is proven. Easy to use. Offers flexible access to course. Learning is self-paced.	Development is expensive. Generic titles are expensive. Delivery system is expensive. Laser vision is not portable. Update is not possible with disc. Early generic titles are poor quality.
CD-Based Technologies	Relatively cheap development. Minimum multimedia on existing PCs. Delivery platform is cheap. Compact discs are cheap to produce. Easy to update video files. Variety of tools to construct materials. Platform no longer dedicated to training function. Digital files can be sent via networks.	Limited generic titles available. No single standard of motion video. Motion video quality worse than analogue. Compact discs are still evolving. Developing interactive programs requires skills.
Communication Technologies	Immediate contact with learners and tutor. Maximum use of existing hardware. Support when suits individuals. Learning assisted at appropriate times. Appropriate technology skills are developed which are transferable.	High costs of set up. Skills needed for technology and handling learners. Technology can be unreliable. Standards needed since it is in its infancy.

Table 2.2: A comparative summary of educational media (adapted from Faint, 1994).

2.6 Evaluation of CAI - The Road to Better Software

Ransdell (1993) echoes a number of researchers when she argues the need for some sort of methodical appraisal of educational software is long overdue. She states that though "Educational software is proliferating at a tremendous rate. Unfortunately, the systematic evaluation of software's effects on learning is not growing nearly as rapidly." Indeed, there are now over 10,000 pieces of educational software available and Anderson et al. (1985) drew on two studies to support their claim that, "A large fraction of CAI software is of low quality and accounts for much of the teacher disenchantment with the computer". The significance of evaluating computer assisted learning therefore cannot be overstated and throughout this work we will return to this theme again and again. One researcher, Castellán (1993), notes that there is a critical need for evaluation to be carried out effectively since he thinks "we all share the vision that information technology has the power to transform teaching and learning, but that it cannot do so without evaluation". Although it is relatively easy to propose what a successful CAL system should possess, it has been more difficult to develop suitable methods or measures to check the effectiveness of such systems and in this section we shall examine some of the reasons for this. Firstly, we can illustrate how easy it is to provide guidelines for effective educational software by examining the following table:-

Effective CAI

- Specifies instructional objectives
- Matches learner characteristics
- Maximizes interaction
- Adapts to the needs of individual students
- Maintains student interest
- Approaches the learner positively
- Provides a variety of feedback
- Fits the instructional environment
- Evaluates performance appropriately
- Uses the computer's resources wisely
- Is based on principles of instructional design
- Has been evaluated thoroughly

Table 2.3: Characteristics of Effective CAI (page 23, Hannafin and Peck, 1988).

These overall goals seem praiseworthy, but they lack clear prescriptive design directions. For instance, an important deficiency of CAL highlighted was the students' level of competency that could not be assessed. Part of the problem with this is the lack of sophistication in the knowledge base (a matter which is being tackled by intelligent tutoring systems), as well knowing what is actually happening inside a particular student's head, however a requirement of educational software is to provide specific feedback on student errors, as well as recording and presenting information to the student about his performance as he proceeds through the task. This is in addition to the information about his overall success and diagnostic information relating to his weaknesses which would benefit from more attention. Supporting evidence of this is provided by Stones (1983) from his reference to two studies, in which human teachers gave different types of feedback which affected their pupils subsequent learning. Pupils who scored the lowest were the ones who were given global encouragement, while detailed feedback resulted a better performance. Whereas *provide a variety of feedback* maybe a good goal, what is really needed from an evaluation is clear design improvement targets. Carrying this out as well as making sure it is a tight research design is a very demanding task.

Bell (1985) relates the experiences of a Coventry-based project where microcomputers were used to provide training for young adults (who had just left school) on the Youth Opportunities Programme (a government apprenticeship scheme). The study represents an early example of evaluation of computer based learning (CBL) in an applied context. One of the project's aims was to evaluate CBL against other and more traditional methods in providing cost-effective training. She highlights that a model was established for the use of CBL and this model was then used to assess the impact of this project. The results showed that CBL provides an ideal way of individualising learning. Bell also reports that "In general students appeared to find the computer an acceptable medium through which to undertake activities about which they were normally reticent". By the introduction of CBL, teaching staff were relieved from the burden of administration and were able to give more concentration to the tuition required. There were also benefits for individual students, some of which who were initially afraid to use a computer became confident after experiencing the frequent and positive feedback and the students as a whole "responded particularly well to the privacy and individual nature of the learning

experience. The immediate and positive feedback given, the endless patience of the system and the opportunity for the student to work at his/her own pace combined to make the learning experience rewarding and stimulating".

This study is useful because it illustrates a number of methodological flaws for this type of evaluative work. Firstly, there were no comparisons provided of the impact of CBL, i.e. before and after effects. Secondly, there were no reported results of learning improvements with CBL, even though CBL was used to deliver tests during the students' lessons. Thirdly, there was a lack of emphasis given to discovering what the students' attitudes were to specific aspects of the courseware, even though general favourability with the system was high. Although it is fair to point out that the courseware was due to be validated by a number of measures, including student performance and attitude, by the project within the framework of a Curriculum Development Group. On the whole, this clearly exemplifies how flimsy a study which attempts to assess and evaluate CAL can be constructed and carried out. While it is true that more contemporary evaluation studies have improved, they still suffer from more intractable problems. In this section we shall indicate a few of the more common obstacles.

2.6.1 Pitfalls in Evaluative Design

Probably the largest source of difficulties originate from the research design. Two authors, Duncan (1993) and Ransdell (1993) have documented some of the main obstacles in achieving a desirable research design for the evaluation of educational technology. Both of them emphasise the choice of a particular method necessarily means the tradeoff between particular issues, such as *internal*, *external* or ecological validity. The difficult choices that are available when selecting the research design will be examined in chapter seven. However, for now we will highlight some of the dangers to evaluative research, particularly those specific to CAL. The following table is a summary of the different types pitfalls suggested by the two authors.

1. THREATS TO INTERNAL VALIDITY (mainly those under student control) Between-group designs:- a) selection bias due to nonrandom assignment of students b) treatment is confounded by variation in students' characteristics c) control of students' characteristics is difficult d) lack of consistency of instruction across sections of a course e) awareness of students that some are receiving CAI while others are not f) uneven rates of student attrition over time g) changes in student performance due to attendance, study habits and variable interest in, and difficulty of individual topics h) verbal reports as data (e.g. questionnaires) are subject to social desirability i) needs of particular class are not anticipated (such as extra-credit work, reviews of student-generated questions and pacing of material with students' progress) Within-group designs:- a) difficulty of equating the subject matter and performance measures over different parts of the course	e) studies conducted outside the classroom setting f) homogeneous samples of college student ability g) limited students or computer resources h) bias of publishing "positive" results i) general instructor as observer biases j) general student as respondent biases 3. PROBLEMS WITH TYPE OF EDUCATIONAL ACTIVITY OR INDEPENDENT VARIABLES a) absence of control group means ambiguity of instructional effect (i.e. computer or other) b) difficulty of equating different educational treatments (e.g. course content, instructors and teaching strategy) c) between-medium comparisons (e.g. CBI and lectures) often contain confounds (a multitude of factors affect classroom learning) d) relative time with each method is significant, as well as order and recency effects and newer media carry a novelty effect 4. PROBLEMS WITH OUTCOME OR DEPENDENT VARIABLES a) performance tests are neither criterion-referenced nor norm-referenced
2. THREATS TO EXTERNAL AND ECOLOGICAL VALIDITY (mainly from the instructor and classroom) a) data collected for computer based experiments with little regard for accuracy or professionalism b) isolating method from general teaching style c) evaluation knowledge gained may be obscured by varying methods to suit students d) studies of short duration	b) test items lack reliability or only have face validity c) when measuring acquired information cognitive style or processing are not taken into account d) students' perceptions are ignored in any questionnaire e) other significant outcomes are absent e.g. how far program stimulates interest in course and career choice

Table 2.4: Potential pitfalls in evaluative research (from Duncan, 1993 & Ransdell, 1993).

Ransdell (1993) describes how the evaluator (in common with other researchers) is caught between the dilemma of attempting to achieve internal validity at the expense of external validity. While the desire to increase the internal validity of a study necessitates tighter control (such as ensuring narrow definitions of independent and dependent variables), this tends to limit external validity (i.e. generalisability to other populations, especially to natural contexts) by virtue of its regulation of extraneous factors. Consequently this issue of control versus generalisability is central to an understanding of the design obstacles which afflict evaluative research. As we shall see, doing research in a classroom setting is an extremely complex task and means that the evaluator has to overcome a lot of these type of hurdles. In addition, quite often this type of research means having to balance the role between evaluator and instructor. While instructors wish their students to learn as much as possible in their course, the evaluator wants to know what impact new methods might have on this learning. As Ransdell points out that evaluation of methods made while one is evaluating knowledge maybe difficult because "research designs must remain fixed over time and students' needs may necessitate changes in design". It is worth bearing these conflicts in mind as we depict a few of the threats to validity in evaluation studies.

One set of threats to internal validity arise from the awareness of students that some of them are receiving CAI while others are not. Duncan (1993) suggests this potentially could lead to resentful demoralisation or even compensatory rivalry between groups. It can certainly pose an ethical dilemma for an instructor who is faced with withholding potentially beneficial programs from some students, particularly if the computer component is tied to course performance and therefore course grade. Within-group designs, are chosen for their increased power and control over student differences, can also have a number of problems resulting from comparisons of segments with CAI and those segments without CAI. Primarily, the predicament is of equating the subject matter and performance over different parts of the course. Indeed Ransdell asserts that any changes other than a simple addition of software are likely to jeopardize internal validity, she gives the example of changes in course syllabus. The result for the evaluator is that it will make interpretation difficult. Ransdell explains why this is the case, "Learning is necessarily a function of many factors, some of which are under the instructor's control, and others of

which are under the student's". She concludes by saying that many researchers simply claim that experimental research conditions are not attainable in evaluation contexts.

Threats which affect the external validity revolve around how naturalistic the setting is and whether it is possible to generalise from that study. Clearly studies of shorter duration (covering less than an entire course) and those conducted outside the classroom are less generalisable. Moreover evaluators are circumscribed by the situation they find themselves in for example, they are limited by the sample of students they have to teach, as well as computer resources they have available and classroom equipment, which often dictates where software can be tested. Duncan reminds us that, as with other types of human research, the context itself can prompt certain types of uncommon behaviour. From her personal experience, using a number of different types of CAI over 3 years, she found students approached the prospect of using the computer lab with little enthusiasm. Ransdell, however, refers to the fact that students are often overly motivated to get positive results, so they will attempt to excel in the course (in a similar fashion to the well-known Hawthorne effect). Clearly the solution must be to include this set of factors in the design, as Duncan declares "If the behavior of students who are using educational software affects experimental outcomes, such behavior should be the subject of research". Finally, investigators can also affect learning either knowingly or unknowingly. Not only are evaluators motivated to get positive results for publication, but also to justify the claims of the software.

Ransdell relates that most evaluation studies are comparisons of CBL with traditional forms of teaching, such as lectures. Some researchers add that this can threaten internal validity because one should be comparing the messages (i.e. the symbols that carry meaning) rather than solely comparing the media. So that studies comparing superior learning with computer software often contain confounds; for example, merely extra study time with the computer (rather than the unique characteristics of software) may improve learning. Another cause of confusion is that a newer media is known to cause a novelty effect. Duncan underlines this by saying that "The selection of a proper activity for comparison is difficult, for two reasons: (1) The complexity of educational software makes it difficult for researchers to specify the essential characteristics of the educational activity

it provides; the independent variable is thus not clearly defined. (2) As Reeves (1986) points out, the differences and commonalities between new and traditional (e.g., lectures) types of instruction are often ignored or not understood". As a consequence one should be wary of drawing too many firm conclusions from these types of comparative studies.

There is also evidence that computer instruction should be considered in combination with other instructor factors, as a result there is ample reason to refrain from the simple comparison between newer forms of instruction and the traditional ones as to which one is better and instead one should ask *how* is it better. The purpose of the current research is to tease out this very question of *why* CAL is beneficial, with the intention of passing this information onto others. This line of reasoning was supported by a study carried out by Ransdell (1993) using a questionnaire, where the results suggested that the effect of the instructional method may contribute to classroom learning at least as much as any specific effects of the software. Apart from the typical dangers to the validity of outcome (dependent) measures, such as not using criterion-referenced or norm-referenced performance tests and a lack of reliability, these measures also suffer from a lack of variety. Duncan informs us that a wider variety of outcome measures ought to be employed to determine the merits of educational software. Performance tests alone are inadequate and other dependent measures should be utilised, namely student attitudes toward CBL and the degree to which educational software stimulates interest in the course and career choices. She comments that even the measurement of learners' characteristics is a rare event in evaluation research.

2.6.2 When does the Evaluation end and the Teaching begin?

Castellan (1993) acknowledges that attempting to evaluate instructional technology is a formidable task because "Traditional evaluations are problematic, especially if the use of technology challenges our traditional assumptions about what should be learned." He qualifies this further by pointing to another level of complication, "The effectiveness of instruction depends on an appropriate evaluation of that instruction. Evaluation has several components, perhaps the most important of which is the evaluation of outcomes. As a

result, what students learn becomes an important criterion for evaluation. But how and when to assess learning and comprehension is an important and continuing problem. Do we evaluate just at the time at which the student learns an important concept - say, at the end of a class; or at the end of the course - say, in a final exam; or by considering a portfolio of products that the student develops over a period of years? Clearly, the effective evaluation of instruction is a continuing process, which ranges from queries about immediate comprehension to the assessment of long-term effects". While many of the problems discussed so far are common to applied research (or even research in general), an argument can be made that the assessment of students solving a problem using a computer is rather unique.

Underwood and Underwood (1990) specified three specific reasons why the evaluation of educational software is such a special activity: (a) product versus process evaluation, (b) linear versus non-linear work, and (c) individual versus collaborative work. They found that when assessing any piece of work on a computer they could either use the end-product of learning as a measure of educational outcome or take measurements at fixed intervals throughout the whole process of learning. The choice of which one was used reflected the goal of the investigator (this mirrors the historical debate about summative and formative evaluation which will be discussed later in chapter five). They argue the advantages of the computer should mean assessment ought to be focused on the process rather than the product. In addition they confirmed Castellán's finding that a computer-based curriculum is more likely to be a spiral form, i.e. computers gives us the freedom to return to the work again and again. The concern expressed here was that at present we lack an appreciation of the non-linear basis of computer-based work. Finally, they describe a third challenge created by these computerised classrooms, that of attempting to assessing individual work in groups. They declare this problem has a long history, but is amplified by use of computers. For instance "Instead of asking children to write individual stories we can ask them to produce newspapers and multi-media presentations involving graphics, sound and text". They cite a possible technological solution would be to keep a log of the investigative strategies of the students to monitor on-going process. Clearly an evaluator must be aware of such issues when devising or reporting the effect of an instructional method on learning outcomes.

2.6.3 Instructor versus Evaluator

Applied research nearly always means dealing with conflicting goals, however, what is important to realise in the educational setting is that educational treatments are already being applied and anything which interferes with this likely to prove detrimental to the participants. It is worth bearing in mind that while the evaluator is required to obtain useful results, however, the students must be treated with care and therefore the evaluator is likely to be performing an awkward balancing act. This divide between instructor and evaluator can easily be illustrated by examining the following four articles in the evaluation literature, the first two originate from an educator standpoint and the last two are evaluation studies which involve the manipulation of treatments and measurement of outcomes.

Welsh's (1993) paper provides five main suggestions for the successful implementation of CAI at college level. Firstly, she points out that computer instruction can be time efficient and cost effective when implemented properly within a college course. Its advantage can be to free up teachers from technical duties, such as the preparation of demonstrations and experiments and provide the students with an opportunity to utilise time effectively by working through a computer program at their own pace. Secondly, she emphasises that software should be chosen wisely and teachers should be involved in its evaluation. This means making sure the software is not a textbook transferred to computer screen, but instead contains interactive examples or review questions with immediate feedback. In addition teachers ought to be shown how to develop and evaluate software. Thirdly, she argues against the myth that students are afraid of computers. Indeed many of them report a favourable attitude to computers and at college level had at least some experience with them. She attributes this fear as originating from the mathematically oriented courses, with which they are most often associated, rather than the computer itself.

Fourthly, Welsh warns that the use of computers should not be limited to tutorials, drills, simulations, demonstrations, experiment generators, and statistical analyses. This is because CBI is often as good as the teacher. Studies which report better student

performance with a teacher than with CBI may be overlooking the fact that students pay attention and sit through an entire class when taught traditionally but when faced with a computer screen they have a tendency to "skip through the instructions, examples and discussion of the project and simply work on the test questions". One way to remedy this is to produce a graded exam with interactive practice questions on the computer or use CBI to review the student's understanding of the course material before the exams. Finally, she recommends that investigation of CAI and CBI must continue. While she acknowledges that carrying out comparative research is flawed because of the confoundings involved in separating the method from the medium appear overwhelming. It is also true that "Outlining these factors will help us implement CAI and CBI better and convince others to do the same". Overall, she is convinced that computers can be successfully implemented into the college curriculum by adopting these sensible steps, despite the fact they are not key to empirical evaluation.

Reiser and Kegelmann (1994) focus attention on Welsh's second suggestion, i.e. making sure software is selected wisely and teachers are involved in its evaluation. They point out that educators have been finding it increasingly difficult to choose software which is instructionally effective particularly when so much instructional software is becoming available for use in schools. Therefore many software evaluation organisations (more than 30) have been set up to review and assess the quality of instructional software and share the results of that evaluation, so that educators are able to "make wise choices when they go about selecting software for use in their schools". The purpose of Reiser and Kegelmann's article was to review the software evaluation procedures used by some of these organisations. They make it very clear, they are not interested in the type of evaluation conducted by individual teachers for *themselves and their students*, where the primary aim is to become familiar with the software before use by the students or more rarely to decide whether it should be used by the students at all. This kind of evaluation should be more correctly known as a review. Instead Reiser and Kegelmann are concerned with the type of evaluation carried out for a software evaluation organization "to critique a software package so that *other educators* can make informed decisions as to whether to use the software with their students". Their paper scrutinised this kind of evaluation method.

Reiser and Kegelmann reviewed the evaluation procedures used by 18 organisations who evaluate instructional software and examined more than 30 journal articles on various software evaluation procedures. As a result they managed to answer five main questions about this kind of evaluation. Firstly, they asked "Who does the evaluating?" and found that, in most articles and most evaluation organizations, teachers were recommended or were responsible for rating software. Other individuals suggested as evaluators were subject matter experts, media specialists and school administrators. There were very few models where students were suggested as evaluators. Reiser et al. advocate students should be used as *participants* (i.e. where other evaluators observe the students as they use the software and draw conclusions based upon their observations while the students use the software, the amount students have learnt from the software and from eliciting student opinions about software) rather than students as *evaluators* (i.e. where students review the software using some of the processes and criteria used by other evaluators and then make judgements about the quality of the software they have reviewed). Other issues relating to evaluators were that only a small number of the reviewed material indicated how many individuals should act as evaluators (typically identified as three people) and very few explicitly mentioned whether evaluators received any training in the use of the procedure.

The second question was "What is the nature of the evaluation process?" and the authors found most often it involved the evaluator going through the software program in the same way a student would, while in a few cases evaluators would observe students as they worked their way through the program. Other references suggested a combination of these two approaches. As to how this was carried out, almost all of the evaluation procedures involved using a rating form. The process of this evaluation varied a great deal, from those who rated individual features of a piece of software (which were then either just simply added together to represent an overall subjective impression or differentially weighted before being added together for an overall rating of the software) to those where the program was reviewed holistically based upon a rating the whole program on four or five scales (i.e. independent of the individual features). The rating form used was either a Likert-type scale, a checklist of features, or a combination of both approaches. Reiser et al. raised a third question as "What features are evaluated?" and they reported that a wide variety of software features were examined by each evaluation method, ranging from

4 to more than 300. Moreover they were "struck by the lack of consistent terminology across the descriptions of various evaluation methods". Despite this confusing picture they were able to tease out a few categories of features that reoccurred most frequently in the evaluation methods: content, technical characteristics, documentation, instructional design and learning considerations, the objectives of the software, and the handling of social issues. Although they found several software evaluation methods called for the collection of student attitude data (i.e. what students liked or disliked about the software and the reasons why), very few methods required evaluators to discover how much students learn from a particular piece of software (even though a few stressed the importance of collecting student performance data). While those that did collect performance data, often accompanied it by the use of other software evaluation procedures in the research design. The authors reported, one interesting study where the developers were judged on whether they had provided any empirical evidence to demonstrate the software's instructional effectiveness and this constituted an important criterion to judge the software. Overall they state software evaluations often focus on the technical characteristics of software rather than on the instructional effectiveness of the software.

The most important question asked by Reiser and Kegelmann was "How reliable are the evaluations?" and they discovered from several authors that subjective judgements have not shown to be valid indicators of the instructional effectiveness of software. Moreover, others have reported that subjective ratings differ greatly across individuals and groups and have been demonstrated as less reliable than a more objective set of criteria. Furthermore when subjective evaluation techniques are employed, teachers and students often rate software quite differently. Indeed students have been found to be more critical of the instructional software than teachers. One reason for the difference in perceptions may be because different groups tend to consider different features of software programs to be most important. This is supported by evidence that "teachers and software developers tend to value technical aspects of software most highly, whereas those working directly for software evaluation agencies tend to focus on issues of content and instruction".

Finally, the authors questioned "How might the evaluation process be improved?", which focused on the bias of subjective techniques and its sole use to evaluate instructional

software particularly in the light of the previous arguments against this. They pointed out a number of remedies, for instance several authors suggested including student use in order to overcome problems associated with subjective evaluations. Another recommended solution was to observe students as they work through a software program in a classroom as well as in a controlled laboratory setting. While others suggested evaluators should measure what students learn as a result of studying the program as well as observing students. Ideally this means testing students both before and after they complete a software program. As a laudable example Reiser et al. cite a particular study where the model involved "pretesting learners, observing them as they work through the software program, and then posttesting them to identify what they learned". It also included collecting a student attitude data as an important component. This model proved to be more discriminating than a purely subjective technique when it came to investigating which software was more effective with the target learners. The authors conclude by recommending what organisations (responsible for evaluating software) should do to improve their evaluation methods. Students should be participants in the evaluation process, which means they should be observed as they used a software program, asked to share their opinions about the software program and more importantly assessed as how much they have learned from the software. The authors' intention was that these measures will enable software evaluation organisations to identify software that enhances student learning. While their aim was laudable it does not tell us *how effective* the software is and more importantly *how to improve* it.

Guthrie and McPherson (1992) carried out a study comparing two groups of physical education students, one group was required to use a customised CAI in addition to their normal lectures, while the other group only attended the lectures. No significant difference was found between the two instructional treatments, although students who received the CAI scored higher on a competency test than those who didn't receive CAI. Guthrie and McPherson state that "Caution must be used when interpreting the results since failure to detect significant differences in all analysis may have been due to the true differences in the test scores being very small, too large a variability, too small a sample size, or the sensitivity of the dependent measures". They conclude the student attitudinal response towards CAI appeared to be the best indicator of the effectiveness of this technology in

this study. While it was a proficient study, what they failed to consider was the effects of treating the two groups of students differently.

A particularly significant evaluation study was conducted by Thomas and Bostow (1991) in order to determine whether the field of personal counselling could benefit from the use of effective self-instructional programs. Sixty-one students were randomly assigned to four groups and each group experienced a different way of teaching therapy concepts. A week before the study every student was given a paper and pencil pre-test and prior to the research everyone was required to fill in a questionnaire requesting information about their characteristics. During the study, group 1 read an abridged version of a standard text on counselling, group 2 read 300 statements using the therapy concepts, group 3 read the same statements but were asked to fill in the blanks and group 4 read the 300 frames from a computer screen and filled in the missing words by typing in the answers. A post-test was administered immediately following the completion of each student's form of instruction and subsequently they were asked about their specific attitudes regarding their participation in the research.

The findings showed that prior knowledge had no bearing on the results of this study. In addition, improvements in performance on the tests increased in the following order: group 1, group 2, group 3, and group 4. Indeed, they found groups 2, 3 and 4 performed significantly better than did group 1. Thomas and Bostow (1991) state "These results suggest that *overt student interaction* produces the best learning". Furthermore they were able to demonstrate that working with computers appeared to be powerfully motivating for the students. Despite conducting the study for group 4 in an area close to high extraneous noise levels, none of the students complained and instead continued working, failing even to take a break. The follow-up questionnaire backed up this finding and they discovered that group 1 liked their condition least of all and members of group 1 indicated they would have disliked an entire course taught by prose delivery. Thomas and Bostow concluded that "This research confirms the power of automated instruction, especially when a high level of meaningful student interaction is required. Its results are important for computer software tutorial developers". This was a better evaluative study than Bell (1985) for several reasons. Firstly, the study actually measured student performance and

secondly measurements were taken both before and after the instruction. Thirdly, individual differences were taken into account by requesting student characteristics. Fourthly, as we have seen, it was useful to obtain the students' attitudes regarding their experience of the study.

An additional clue to the cultural divide between evaluators and instructors comes from Underwood and Underwood (1990); they discovered that classroom teachers sometimes reject the results of experimental studies in education because the artificial conditions of testing cannot be repeated in the conventional classroom. Underwood et al. lament when they say "The misfortune is that educators are neglecting valuable information: experiments do tell us the specific conditions under which an educational manipulation can be seen to have an effect upon children's thinking". Evaluators and instructors must value each other's work and a balance needs to be struck when carrying out evaluations in education.

2.6.4 Establishing the goals of educational software

We have discussed the pitfalls in evaluative design, problems of measuring educational outcomes with computers, the role of the evaluator (versus instructor), but in order to identify effective educational software and more importantly improve its capability we need to describe the characteristics of computer software that will achieve desirable educational outcomes. Underwood and Underwood (1990) suggest that "The goal of introducing software into the classroom must be to assist with the intellectual development of its user". But which criteria should we look for to meet this goal? Underwood et al. start from the axiom that activities that motivate children also lead to successful learning. Consequently they assert "Classroom software that motivates children will therefore produce successful learning, and will involve children in learning", which begs the question about the nature of involvement and the type of learning which needs to take place. From reviewing current research, they found that software that gave learners control is most likely to be successful (in terms of desirable educational outcomes). In addition the learner needs to be challenged and to have a clear and personally meaningful goal to

aim for. What seemed to be important was that the children needed to be successful, not all the time but often enough to generate a good self-image and raise self-esteem. They comment "This will not be achieved if the child has no idea of how to start the task at hand or indeed if the task cannot be identified. What is a challenge for one child is a daunting uphill struggle for another". Accordingly they contend the style of teaching is important in creating expectations of the correct manner of going about a task, i.e. there should not be a mismatch between the software style and the classroom activities.

How do we know whether these aims for educational software are valid? Well, evaluating educational software is an important task and this type of research must be carried out correctly. It is difficult to carry out any meaningful research in a domain in which the object of evaluation is currently being updated and there are no formal standards. Designers of educational systems should be setting targets of what their software ought to accomplish after being used by a particular user. Moreover, there should be more independent assessors because designers cannot always recognise improvements to be made to their own software and there should be more public acknowledgment when systems are successful. This style of evaluation seems to be the only way to increase the effectiveness and usefulness of such systems, resulting in better software, wider usage and a longer shelf-life.

2.7 Key Points

- Computer assisted instruction (CAI) has long been heralded as one of the greatest innovations in education and has been proliferating at a tremendous rate. Unfortunately, CAI have generally failed to live up to these early expectations and a large fraction is low quality.
- CAI systems have improved in technical complexity but have failed to bridge the transition from research laboratories to the classroom. Development of high quality software takes a huge amount of time, money and expertise.

- Evaluation is a vital activity for keeping the development process on track (i.e. requirements-driven development cycle). It also serves to publicly acknowledge the worth of an instructional system.

- To increase both the effectiveness and usefulness of CAI, so that there is better software, wider usage and a longer shelf-life, we recommend:-

- (i) designers of educational systems should set targets of what their software ought to accomplish after being used by a chosen user;

- (ii) persuade more independent evaluations, since designers cannot always recognise improvements to be made to their own software; and

- (iii) promote more public awareness when operational systems are successful (to serve as exemplars).

- Ensuring a good evaluation design is an awkward balancing act. The choice of an evaluation method necessarily means a trade-off between: (i) internal and external (or ecological) validity; (ii) the role of instructor versus evaluator; and (iii) the special problems associated with assessing the outcome of educational software (Underwood and Underwood, 1990).

- Examination of the evaluation literature provided us with examples of evaluation studies, from which we derived practical guidelines for conducting good evaluations. For instance, Thomas and Bostow (1991) was a much better evaluative study than Bell (1985) because: (i) measurements of student performance were taken, (ii) they were taken before and after instruction, (iii) individual differences were taken into account, and (iv) it was useful to acquire the student's attitude regarding their experience of the study.

CHAPTER THREE

INTELLIGENT TUTORING SYSTEMS

"Individualizing instruction, in the educator's sense, means injecting humor into a lesson when a student seems to need it, and quickly becoming serious when he is ready to settle down to work; it means thinking of examples that are uniquely relevant to the student's previous experience and offering them at just the right time; it means feeling concerned over whether or not a student is progressing, and communicating that concern in a way that will be helpful; it means offering appropriate praise, not just because positive reinforcers strengthen response tendencies, but because the student's performance is deserving of human admiration; it means, in short, responding *as an individual to an individual*." [Jackson, 1966] (1;9;18;23;8;9;12;19)

3.1 What are Intelligent Tutoring Systems?

A good tutor, as the above quotation illustrates, understands and responds to their student's special needs. From its earliest times, the computer has been viewed as capable of providing such instruction. Intelligent tutoring systems (ITSs) can be thought of as attempts to produce computer behaviour which, if performed by a human, would be described as *good teaching* (Nwana, 1990). In essence, like traditional computer assisted instruction (i.e. CAI), they are computer programs designed to present educational material to the student. However unlike CAI they incorporate techniques from artificial intelligence (AI). Indeed they are colloquially known as intelligent computer assisted instruction (ICAI), denoting their attempt to model the expert human tutor more closely than other forms of CAI. This is usually epitomized as a goal of individualisation of teaching.

Goodyear (1991) suggests there are two main ways of enabling such a system to adapt its teaching to the individual needs of different learners "Using conventional programming techniques, the system builders must prespecify all available routes through the space of teaching possibilities. Every test, every decision, every branch to some remedial material

and every exposition must be written in advance. For teaching with any significant complexity, the combinatorial problem is enough to render the design, comprehensibility, and subsequent maintenance of the system highly problematic. This suggests a second approach, which is to enable the system to make decisions, to reason during the teaching interaction. AI provides an expanding set of tools for simulating reasoning - for carrying out inference - in computer systems. A prerequisite for such reasoning is that the system must have available to it an explicit representation of the knowledge it needs (in its domain of expertise). By *explicit*, we mean available in a form on which the algorithm of inference can work". (20;3;15;23;1;18;4;12) He argues this problem of formalizing the relevant knowledge forms a large part of the effort in creating artificially intelligent systems. So in the case of artificially intelligent teaching systems Goodyear views the task as conventionally revolving "around the formalization of three (or four) sets of knowledge". The four types of knowledge are: *what subject* to teach, *who is* to be *taught*, *how to teach* and *how to display* a particular set of information. Nwana (1993) has referred to the first three of these as the traditional trinity and they are formally represented as components of ITSs.

Although the architecture of intelligent tutoring systems varies considerably, Nwana (1990) has cited the general form of an ITS as consisting of four basic components mentioned earlier, that is:

- (1) **The expert *knowledge* module** - which comprises of the facts and rules of the particular domain to be conveyed to the student i.e. the knowledge of the experts.
- (2) **The student *model* module** - which refers to the dynamic representation of the emerging knowledge and skill of the student.
- (3) **The *tutoring* module** - the part that designs and regulates instructional interactions with the student (which can also be referred as a teaching strategy or pedagogic module).
- (4) **The user *interface* module** - which is the communicating component of the ITS which controls interaction between the student and the system.

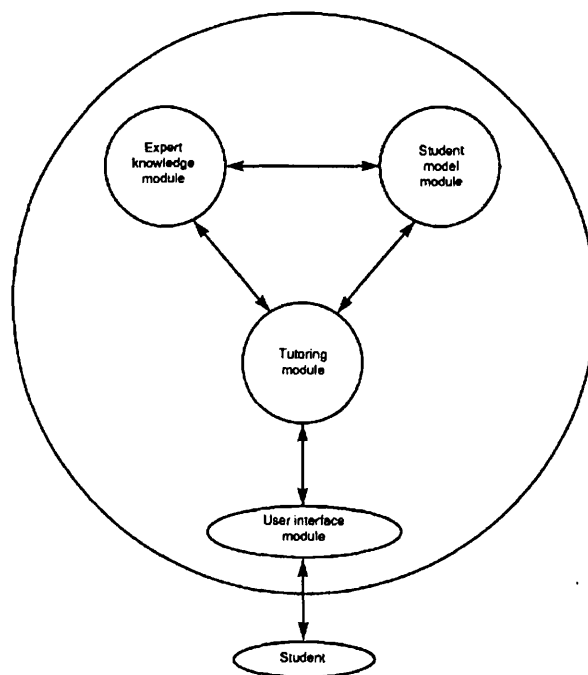


Figure 3.1: General ITS architecture (from Nwana, 1990)

In this chapter we present an introduction to the way artificial intelligence is being applied to education. The aim is to explain some of the background to the ITS research work and a number of the issues surrounding this field. The particular issue chosen to focus on is the most significant, namely *evaluation*.

3.2 History of ITSs

A landmark paper by Jamie Carbonell (1970) proposed a new approach to computer-assisted instruction (CAI) and in doing so he re-defined the field. He wanted to include artificial intelligence into CAI systems, indeed "He dreamed of a system which had a database of knowledge about a subject matter and general information about language and principles of tutorial instruction" (Nwana, 1990). Later, Self (1974) argued that a computer tutorial program should not only have a representation of *what* is being taught, but also *who* is being taught and *how* they should be taught. This separated the field into the traditional CAI system (which presented portions of the curriculum to the students in a

fixed way) and the new area called intelligent tutoring systems (ITSs). Carbonell stressed the need to anticipate all possible interactions between the system and the student. The transition from CAI to AI-based systems, shifted emphasis from constructing a curriculum of exercises to representing the knowledge necessary to solve a restricted set of problems. This was reflected in the acute difference of ITSs which borrow more from the field of cognitive psychology as opposed to CAI which had a behavioural inheritance. Carbonell constructed an early version of such an ITS called SCHOLAR, but he died before it was fully realized. Carbonell's dream, however, did live on through the large impetus created by him for an era of ITS development.

So much research had taken place in intelligent CAI systems that in 1978 two prominent researchers, Derek Sleeman and John Seely Brown, edited a special issue of the International Journal of Man-Machine Studies which attempted to summarize the research in this area. The title of their popular book, which was published in 1982, gave the field its commonly used name, i.e. *Intelligent Tutoring Systems*. However, early intelligent tutoring systems were still viewed as improvements on the CAI systems. In fact Yazdani (1987) states the ICAI systems of the 1980s were "attempting to deal with the shortcomings of generative systems". ITSs were based on providing the computer with an explicit description of the subject matter to be taught and the methods of teaching. As a consequence, course material became independent of teaching procedures so that problems and remedial comments could be generated differently for each student. In addition, the teaching method could be varied without having to change the description of the domain. Another significant advance on the CAI was that they attempt to form an assessment of the individual learner and use this to tailor the learning experience to that particular individual. In fact Yazdani (1986) expressed this clearly when he said "The notion that the tutoring program itself can solve the problem which it is setting for the student, and in a way similar to that of the student, is the basis of a large number of ITSs".

Having said that, ITSs demonstrated a number of characteristics over and above the traditional form of CAI. Carbonell had specified some of these characteristics as: mixed-initiative dialogue, semantic (knowledge) networks, student models, diagnostic error rules and natural language. Mixed-initiative dialogue refers to the capability for the student to

ask a question and hence participate in a two-way interaction with the program. This is in contrast to the typical one-way interaction (presenting question or problems followed by a student response) of traditional CAI. We have already discussed the separate representation of knowledge, such as facts, rules and their relationship. In ITSs these are usually in the form of semantic networks, in contrast to the scriptlike structure of most traditional CAI programs where all the content is organized into screens and branching instructions.

Moreover since the instructional sequence was not predefined as in traditional CAI, it was necessary to have a good idea of what the students knew and what they have already learned. This meant possessing a student model as well as diagnosing the errors made. Consequently this was achieved by identifying the aspects of the knowledge network currently understood by the student. In addition it was necessary to diagnose any mistakes made by the student (i.e. identifying the student's present level of understanding - for example, in terms of misconceptions, overgeneralisations, and missing information). The latter is done by including both a general content-independent error detection mechanism and a content-specific error mechanism in the subject domain. Another characteristic of ITSs was natural-language interaction. Obviously if the program can understand natural language input, it is likely to improve communication between the student and the system.

While there have been many new developments since Carbonell's classic paper, his five characteristics remain some of the important concepts of ITSs. One significant change is that the understanding of knowledge networks has gone beyond strictly language-based (i.e. semantic) representations. In fact, the importance of natural language has diminished considerably in ITSs (and AI generally). This is because it is possible to implement intelligent systems using structured command or menu selection structures. This shift corresponds to a greater interest in the inferencing mechanisms associated with tutoring than with understanding language.

Nwana (1990) relates that by the early 1980s researchers in this dynamic new field had split into two camps. The first became known as the *exploration environments* group because it aimed to develop environments which encourage discovery learning (i.e.

learning by doing). Seymour Papert is the best known advocate of this school. Papert considers that the presence of computers will lead to learning without schools and consequently "schools as we know them today have no place in the future" (Nwana, 1990). Clearly this is a radical approach and has been greeted with criticism by some in the education field and therefore members of this school are often referred to as revolutionaries. A number of these systems have been created and are characterized as having no explicit representation of goals, state or performance and no capacity to direct student behaviour. Examples include Papert's Turtle system and Alan Kay's Smalltalk system. Kimball (1982) cites the common accusation against the environmental approach as providing systems with "no facilities for detecting, rescuing, or motivating the student who was in trouble". The second and the much larger of the camps is the *intelligent tutoring group*. Nwana (1990) distinguishes this group as supporting a gradual improvement in the present quality of education using AI techniques. He also indicates "They advocate a paradigm where the computer acts a tutor, i.e. students largely learn by being told" and therefore they are reformists since they favour evolutionary change.

Chuen (1989) enumerated the five basic types of ITS which have been depicted in the literature. The *Mixed Initiative Dialogue* was the original ITS paradigm advocated by Carbonell. This type of program engages the student in a two-way conversation in which both the student and machine can initiate a question and attempts to teach the student using the Socratic method. However the trend has been away from this kind of tutor *towards coaches* and *diagnostic* tutors. The former are best suited for problem solving situations (e.g. simulations and games), and these programs act behind the scenes and leave the users with complete control of their learning activity. The latter compares the students' behaviour with that of an expert for the problem domain involved, so that advice is given when discrepancies arise (e.g. misconceptions and bugs). *Microworlds* are computational tools that simulate the domain under the study. As we have seen, this type of program allows students complete freedom. Since no active instruction is advocated, students are encouraged to formulate their own concepts of the domain. Finally there is the *Articulate Expert System*, which is an expert system that can explain student actions and decisions and learning takes place by the student asking questions.

As we have stated previously, at present there does not seem to be an agreement on ITSs architecture. However a few individuals like Yazdani (1986) have suggested a simple classification, where AI applications fall between two ends of a spectrum. Yazdani recommends that the three different proposals of Anderson, Boyle and Reiser (1985), Hartley and Sleeman (1973), and O'Shea et al. (1984) can be viewed as points on a spectrum, where learning (or open-ended problem solving) environments are at one end and traditional CAI systems are at the other.

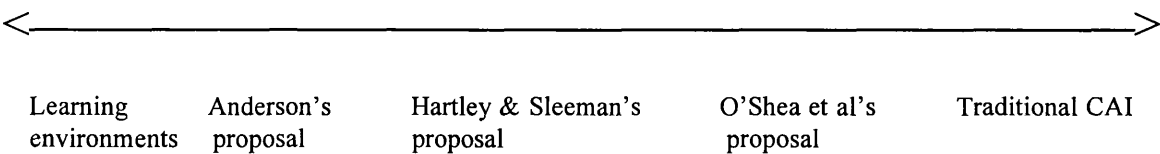


Figure 3.2: Continuum of CAI types (adapted from Yazdani, 1986).

Yazdani indicates that the position of each architecture, on the above spectrum, reflects the nature of the expertise that is taught. So that on the one hand, highly abstract and general concepts, e.g. model building is better taught within an exploratory learning environment by the construction of an appropriate computer-based microworld, whereas tasks that become more concrete and specific are best achieved within the traditional CAI domain.

Looking back on this field, it has clearly had a tremendous period of activity ever since SCHOLAR. Predominately it has been able to demonstrate that innovative ideas from AI can be applied to education, in fact the tutor SOPHIE predates most of the work carried out in AI. Moreover a sufficient number of ITSs have been developed for almost any subject area or training domain to establish the potential of ITSs. An illustration of this productive nature is provided in Table 3.1, which details the sheer variety of ITSs reported in the literature (although we must caution most of them do a lot less than it is claimed). As for the current state of ITS development, Goodyear (1991) reports this is more difficult to determine. However he does say, "It is clearly more advanced in those domains which seem to lend themselves more readily to formalization (some areas of physics, elementary mathematics, computer programming, equipment control and maintenance, for example). It is also more advanced where system builders have adopted pedagogical models which

could be described as instructional, goal-directed, system-led or content-driven rather than models which favor learner-directed exploration, negotiation of purposes, or the acquisition of higher-level cognitive skills". He highlights a number other features about the contemporary and future ITS scene. Firstly more learner-based models are being explored, distancing themselves from the mainstream intelligent teaching systems. Secondly progress is being made through at least two styles of working:-

a) since the field is still too young to justify tackling complete ITS, the focus is on studying BITS (Bits of an ITS), i.e. domain knowledge, student model (or learner model), teaching knowledge and explicit representation of rules for dialogue are each sufficiently rich and elusive to warrant independent study,

b) while perceive that knowing about the BITS will not tell us how to put them together or the relation of each unit to the whole, consequently the central problem must be to study the integration or interplay of each of the BITS.

Thirdly there are a number of strategic sites for research: problems of learner modelling (especially, methods for enriching or extending the content and utility of learner models), the implementation of less didactic teaching styles (e.g., through collaborative learning or guided exploration), improved methods for the acquisition and representation of pedagogical and domain knowledge, and the exploitation of better methods for learner-computer communication (e.g., through graphic interfaces and more sophisticated dialogue modelling).

ITS	Domain	Reference
ACE/PSM	NMR spectra interpretation	Sleeman (1975)
ATDSE	Basic subtraction	Attisha & Yazdani (1983)
ARITHMEKIT	Basic subtraction	Brown (1983)
ALGEBRALAND	Algebraic proofs	Brown (1985)
BIP-I/BIP-II	Basic programming	Barr et al. (1976)
BLOCKS Tutor	Troubleshooting in BLOCKS world	Brown & Burton (1978b)
BRIDGE	Programming	Bonar (1985)
BUGGY	Basic subtraction	Brown & Burton (1978a)
DEBUGGY	Basic subtraction	Burton (1982)
EDSMB	Basic multiplication	Attisha & Yazdani (1984)
EUROHELP	UNIX mail	Breuker (1987)
EXCHECK	Basic logic	Blaine (1982)
FGA	Basic French grammar	Barchan et al. (1986)
FITS	Basic fractions addition	Nwana (1990)
FLOW Tutor	FLOW computer language	Genter (1977)
GEOMETRY Tutor	Geometry proofs	Anderson et al. (1985a)
GERMAN Tutor	Basic German	Weischedel et al. (1978)
GUIDON I/II	Basic medical diagnosis	Clancey (1987)
INTEGRATION Tutor	Basic integral calculus	Kimball (1982)
LISP Tutor	Lisp programming	Anderson & Reiser (1985)
LMS	Basic algebra	Sleeman & Smith (1981)
MACSYMA Advisor	Use of MACSYMA	Genesereth (1982)
MALT	Basic machine language programming	Koffman & Blount, 1975
MENO-Tutor	Basic Pascal programming	Woolf & McDonald (1984)
METEOROLOGY ITS	Basic meteorology	Brown et al. (1973)
NEOMYCIN	Medical diagnosis	Clancey & Letsinger (1981)
PIXIE	Basic algebra	Sleeman (1987)
PROUST	Pascal programming	Soloway & Johnson (1984)
QUADRATIC Tutor	Quadratic equations	O'Shea (1982)
QUEST	Basic electrics	White & Frederiksen (1985)
SCENT - 3 Advisor	Lisp programming	McCalla et al. (1988)
SCHOLAR	South American geographical facts	Carbonell (1970)
SIERRA	Learning basic arithmetic procedures	Vanlehn (1987)
SOPHIA I/II/III	Basic electronic troubleshooting	Brown et al. (1982)
SPADE	Basic LOGO programming	Goldstein & Miller (1976)
SPIRIT	Probability theory	Barzilay (1985)
STEAMER	Marine steam propulsion plant	Hollan et al. (1984)
TALUS	Basic Lisp programming	Murray (1987)
THEVENIN	Basic electrical circuits	Joobbani & Talukdar (1985)
TUTOR	British highway code	Davies et al. (1985)
WEST	Basic arithmetic skills	Brown & Burton (1978b)
WHY	Basic meteorology	Collins & Stevens (1982)
WUSOR	Maze game skills	Goldstein (1982)

Table 3.1: A reasonably comprehensive list of ITSs and environments (from Nwana, 1990).

3.3 Examples of Classic ITSs

In the following section we present two different types of intelligent tutoring systems. The reason for highlighting these systems is twofold, firstly because they provide good exemplars of landmark changes in ITS design and secondly they give a chance to comment on benefits and shortcomings of each approach to ITS construction.

3.3.1 SCHOLAR

Since SCHOLAR was the first ITS to be constructed it automatically earned Jamie Carbonell a place in the history of ITS. It was designed to teach the facts about geography of South America. O'Shea and Self (1983) provide the type of dialogue typical of SCHOLAR (where the normal text is the computer response and the italicised text is the human response).

.

.

.

Use one of the following:

Sucre

French

Argentina

Spanish

to answer the question:

What is the language in Chile?

Spanich

You should have written 'Spanish'.

Very good.

The capital in Chile is Santiago.

Correct or incorrect?

Correct

That's right.

Approx what is the area of Brazil?

2500000 square miles

Wrong.

Please indicate if the following statement is correct or incorrect:

The area of Paraguay is approx 47432 square miles.

What is the area of Brazil?

The area of Brazil is approx 3600000 square miles.

Now answer the question you didn't answer before.

False

Rephrase your statement. I can't understand it.

Incorrect

Very good.

The area of Paraguay is approx 157000 square miles.
The language in Brazil is French.
Right or wrong?
Wrong
You are correct.
The language in Brazil is Portuguese.
Is it true that the capital of Brazil is Brazilia?
Yes
That's fine.
Approx what is the population in Brazil?
60000000
Approximately true, but a more exact value is 70000000.
What is Brazil?
Tell me something about Peru
Wait a minute. I'm checking my files.
Let's proceed now.
Peru is a country.
It is in South America.
The population is approx 7000000 people.
The capital is Lima.
Now answer the question you didn't answer before.
q/a
Do you want to change the general mode of the interaction to q/a?
Please answer y or n.
y
Ok, Mr Jones. Let's go ahead.
Please, tell me more about Peru
The area of Peru is approx 480000 square miles.
The language is Spanish.
How many square miles are there in Brazil?
The area is approx 3600000 square miles.
What is the principal religion in Uruguay?
The principal religion in Uruguay is Catholicism.
The principal religion in Peru is-
I'm afraid I can't answer that.
Is it true that the area of Uruguay is 70000?
That's about right. The exact answer is 79000.
Could you please indicate the principal rivers in Argentina?
I don't understand your word 'indicate'. I'll give you another chance.
The principal rivers in Argentina are-
The principal rivers in Argentina are the Rio de la Plate, the Parana, the Uruguay river, the Salado, the Paraguay river, and the Pilcomayo.
Mixinit
Do you want to change the general mode of the interaction to mixinit?
Please answer y or n.
y
Ok, Mr Jones. Let's go ahead.
Where is Brazil?
.
.
.

Figure 3.3: A dialogue with SCHOLAR (from Carbonell, 1970).

O'Shea et al. (1983) observe that "the student responds in English, sometimes with short answers of the kind we have seen with branching programs, but sometimes with apparently unrestricted sentences, for example, to ask a question. Since the questions cannot be anticipated by SCHOLAR, it has to be able to analyse them in order to determine what information it is that the student requires. In fact, SCHOLAR's language processing capabilities were rather limited,". SCHOLAR was the first in a long line of mixed-initiative tutors, designed so that both the student and system could take the initiative in the conversation and ask questions. Therefore, with reference to the dialogue shown, the student typing "q/a" or "mixinit" allowed them to change the mode of interaction and this meant the student was able to ask questions at the time they occur to them.

Clearly the SCHOLAR system does not find answers to questions by searching a long list of facts about South America. O'Shea and Self points out that "For one thing, it would take too long. For another, the particular fact required need not be stored at all. For example, the two facts 'Peru is in South America' and 'The capital of Peru is Lima', plus a little common sense, should be sufficient to enable the question 'Is Lima in South America?' to be answered". The answer is that SCHOLAR represents knowledge as a semantic network - where an element of this network is related systematically to other elements in the network. Consequently, O'Shea et al. state "To answer a specific question, such as 'What is the latitude of Argentina?', SCHOLAR must find a node ('Argentina') in the network, and then find the relevant relationship (to 'latitude'). For South American geography it is not very difficult to organise the network so that a systematic search will locate required nodes sufficiently quickly".

However, Nwana (1990) indicates "the real power of this representation schema comes by recognizing that it is possible to answer questions for which answers are not stored. This automatically relieves the system of the memory problems encountered in anticipating and storing all solutions by traditional CAI systems. For example, it is not necessary to store in the semantic network that 'Lima is in South America' provided that the program which interprets the network can make the relevant inference. In other words, the program must know about the attributes concerned, e.g. 'location' and 'capital', and in particular, that

if x is capital of y and y is located in z then x is in z: this is a rule of inference". (25;4;15;7;19;12;9;11) A shortcoming of this approach is that since there is no teaching strategy, "the questions follow certain stereotypes, with particular topics being selected on an essentially random basis.... SCHOLAR seems to flit from one thing to another, and there is no systematic attempt to develop the student's knowledge in any particular direction" (O'Shea et al., 1983).

Carbonell use of the semantic network emphasised the first attempt to develop a tutorial system to teach "the same way human teachers do". Despite a ground breaking use of natural language based on knowledge of a specific subject at hand and of more general, knowledge it was a system which had narrow requirements. For instance, even though semantic networks are to some extent general-purpose representational schemes and therefore other kinds of knowledge could be expressed in them. However, O'Shea says "SCHOLAR could be adapted, with increasing difficulty, to teach, say, the geography of Africa, the anatomy of the circulatory system, or aspects of analytic geometry, mainly by changing the contents of the semantic network and not the program itself.

The difficulty increases as the subjects become less about facts and more about procedures. To teach the geography of Africa we would need to change only the entries in our semantic network, the structure of the network and the rules of inference remaining largely unchanged. But, as we are sure the reader has noticed, SCHOLAR only deals with a small part of what is normally covered by 'geography': it cannot, for example, explain processes (e.g. how an ox-bow lake is formed) or interactions between systems (e.g. wind, sea and cloud). However, the fact that some of SCHOLAR, at least, could be carried over to a new topic is an improvement on an author language program to cover the same material. This would need virtually complete rewriting, since all the messages to the student would need changing". Here we have hit upon a limitation of such teaching systems, i.e. the way the knowledge is represented *determines* what kind of subject and student understanding the program is capable of.

Wyer (1987) indicates a limitation which follows on from what has just been said, "SCHOLAR was significant for ICAI in that it made some attempt to utilize restricted natural language for both input and output, but it did not meet another requirement set

forward for intelligent CAI systems: it makes no attempt to evaluate the student's incorrect responses or to use a diagnosis of what may be wrong, to help the student overcome misunderstandings or misconceptions. This touches on one of the drawbacks of much frame-oriented CAI, and it perhaps indicative of a certain narrowness of perspective that behavioural-based teaching methods can have, that is, an overriding concern for successful outcome while ignoring the enormous issue of how one gets there". (5;16;18;15;6;5;19;19)

Finally, O'Shea et al. (1983) spell out two additional problems with such dialogue systems and sum up the contribution of SCHOLAR, "First, a program may take a considerable time to understand an English sentence and to compute an appropriate response, and this may well reduce the effectiveness of the tutorial interaction. Secondly, writing a program like SCHOLAR to simulate a human tutor is not a trivial task: it is not one to be tackled by those teachers for whom author languages were intended, for it requires man-years of expert programming.

None the less, systems like SCHOLAR do demonstrate that there is at least a possibility of achieving the goal of developing programs able to engage in meaningful tutorial dialogues, for which the venerable dialogues of Socrates have served as something of an ideal and inspiration. But this goal will not be achieved easily if at all because it requires solutions to problems which we know to be difficult from studies in artificial intelligence,"

3.3.2 GUIDON

GUIDON was a program developed by William Clancey and his colleagues at Stanford University. It represents a milestone in the ITS field because it was the first tutor built to work with an already existing expert system (i.e. MYCIN). MYCIN is a well-known (rule-based) expert system, which acts as a medical consultation system for diagnosing infectious diseases and providing antibiotic therapy. The principle objective of GUIDON was to convert MYCIN into an effective instructional tool. Consequently GUIDON combined the subject matter knowledge of the consultation system with teaching knowledge, while keeping the two distinct.

GUIDON was strongly influenced by previous tutors, namely SCHOLAR, SOPHIE, WEST and WUMPUS. Clancey (1987) also acknowledged that his progress had been helped by previous work (as we shall see something quite rare in this new field), "We sometimes joke about the futility of progress in a field in which scientists stand on each other's toes, rather than building upon past work. In developing GUIDON, I was privileged in two ways: First, I was able to build my work upon the MYCIN consultation program, to which more than 20 people contributed their programming and medical expertise over more than 5 years (1972-1979); second, I was advised by some of the most capable researchers in my field". Another key feature of this project was that the funding came from US armed forces. In fact, quite a number of ITSs built in the USA have been financed by the army or navy.

Clancey chose to simulate the dialogue of a consultant talking to another physician about a patient and aimed to get the student to become more proficient with actual problems solving, i.e. using the case-method approach. As Yazdani (1987) explains, "GUIDON attempts to transfer expertise to the students exclusively through case dialogues where a sick patient (the 'case') is described to the student in general terms. The student is asked to play the role of a physician and ask for information which he thinks may be relevant to this case. GUIDON compares the student's questions to those which MYCIN would have asked and critiques him on this basis. GUIDON uses a 'closed world' model where all objects, attributes and values that are relevant to a case are determined before the tutorial session begins". It is also pertinent to note that communication between the program and the student was limited to a computer terminal that prints one line at a time, like a teletype (in fact subsequent improvements were to add graphical methods of information). The idea is to use the context of an actual problem to make the student aware of gaps or inconsistencies in their knowledge, and to correct these deficiencies.

GUIDON can provide at any point:

1. A list of all data relevant at that point.
2. Sub-tasks to be performed.

3. Hints on how to go about a problem.
4. A summary of all evidence already discussed.
5. Exposition of a task.
6. Completion of a diagnosis if the student gives up.

Nwana (1990) noted GUIDON's natural language capabilities were far better than SCHOLAR's. Moreover since it was able to systematically compare student behaviour to what MYCIN would do and alert the student to this it was a better teaching program than SCHOLAR. Although Clancey (1987) found expert systems are a good starting point for developing intelligent teaching systems, he affirms that "they are not sufficient in themselves for making knowledge accessible to a student. One can follow MYCIN's reasoning during a consultation by requesting explanations (by asking WHY case data are being sought by the program and HOW goals were or will be achieved). However, this is an inefficient process for learning the contents of the knowledge base. The explanation program is only a passive teacher: It is necessary for the student to ask an exhaustive series of questions, if he is to discover all of the reasoning paths considered by the program. Most students would not have this persistence, so the range of expertise in the knowledge base would be lost to them". In addition Wyer (1987) argues the other important lesson of GUIDON (for AI generally) was that since it was unable provide the explanations for sufficient understanding, multiple representations are required for a good range of explanations.

Clancey (1987) acknowledged the difficulties of building such a tutor, "One of the major research problems is the design of tutorial strategies - when to say something and what to say.

Even if we had an accurate representation of a human expert's knowledge, we would need this additional knowledge about presentation and teaching discourse to construct a teaching program. Furthermore, we might need additional facts that play no part in problem-solving itself, to help the student remember facts by relating them to other things he knows or to

help the student resolve conflict when his prior knowledge differs from the program's point of view". Clancey attempted to improve on GUIDON, and a number more effective tutors were created such as NEOMYCIN and GUIDON 2 (for a more details see Richer and Clancey, 1987). Two final points are worth raising in connection with GUIDON. Firstly, although it took five years to develop it doesn't exist today since it became too burdensome to update for new hardware and software. Secondly, a limited evaluation was carried out with one medical student due to lack of time and resources. On the positive side enough was learned from this single experiment to keep the research group busy for years.

3.4 Conceptual and Practical Issues in Intelligent Tutoring

The primary aim and also the greatest strength of ITSs is to provide one-to-one tutoring. In trying to achieve this we certainly need to take into account a number of factors, a significant one is educational wisdom. However this is not as straightforward as it may seem. All attempts to educate make assumptions about the nature of expertise and the nature of the learner, also contain a set of epistemological assumptions. Sometimes these assumptions are explicit; more often they are implicit. Though ITS researchers have offered a rich variety of accounts in each of these domains there does, nevertheless, seem to be a set of core assumptions which can be challenged. Unfortunately, as we shall see, the nature of intelligent systems development means this does not often happen.

Other issues are equally likely to prove troublesome. ITS might be associated with broader social attitudes to education, and indeed to the nature of humanity, which may prove to be seditious. Issues under this heading would be: expertise taken out from the hands of teachers and placed in the hands of machines; ITS as a "panacea for all the ills of education"; devaluation of communication between human beings, such as discussions between pupils and teachers; possible associations between ITS and anthropomorphic views of computers (i.e. imbuing machines with intelligence and devaluing biological intelligence); the links between AI and Cognitive Psychology may lead us to devalue aspects such as open-ended problem solving, creative writing, and musical composition;

and the view that there is a quick technological fix for every conceptual problem. In this section we shall touch briefly on other issues in order to illuminate the nature of the ITS beast.

3.4.1 Importance of ITS research

Given the complexity of individualised teaching, why do researchers bother to produce such advanced computer-based tutors? Essentially the answer to this boils down to two main reasons: research needs and practical needs.

The field of intelligent tutoring is an increasingly vibrant field and as Nwana (1990) points out "it provides an excellent test-bed for various theories from cognitive psychologists, AI scientists and educational theorists. For example, a primary reason why the famous Carnegie Mellon psychologist John Anderson came into the area, was to test out his various theories of learning (Anderson, 1987). Hence, the design of an ITS will contribute to the discovery of more accurate theories of cognition (Burns and Capps, 1988)". This is both an advantage and a considerable challenge. For as Ross (1987), a field worker, puts it "Education is one of the most demanding fields to which AI might be applied. It offers scope to the AI researcher for work in all the basic aspects of his subject: how to represent knowledge, how to reason using that knowledge, and how to control the interplay between these. It also seems to offer a chance to apply existing ideas". Indeed, ITSs have been designed to represent both the concepts to be taught and how a student might learn those concepts. Consequently they have a great potential to influence both education and computer science and have encouraged some researchers to learn about and refine important issues such as: what is the nature of knowledge, what precisely is learning, how can an individual student be helped to learn, and what styles of teaching interactions are effective when?

ITSs also have practical benefits by providing solutions not possible by alternative methods. Nwana (1990) has echoes these sentiments when saying that "On the more applied level, there are a number of useful results which can be achieved using ITSs

which cannot be achieved with human tutors for economic and social reasons (Elsom-Cook, 1987). A primary advantage of ITSs is the possibility for providing one-to-one tutoring. There is a consensus on the view that individual tuition, tailored to the needs of the student, is the most effective form of educational interaction, at least for most domains". What he is referring to is that ITSs have a good track record in instances where money poses no problems for development, students are highly motivated, where domains are heavily factually based and where these facts can be accredited in quantity. Further reasons are provided by Nwana: "Since our educational systems have, of necessity, become geared towards group teaching, many of the advantages of one-to-one tutoring have been lost. ITSs can provide such tuition without necessarily losing the advantages of the group teaching environment (e.g. by providing one ITS per student in a class), thereby getting the best of both worlds". A different reason, for deciding on ITSs is that they embody a *strong* teaching system approach, i.e. that is able to solve the same problem it presents to the student, also it is able to recognize errors, record missing or extraneous components of an answer, and correct errors. Moreover, unlike a *weak* system, it can solve the problem along with the students and assist them when they get "stuck" during a partial solution.

3.4.2 Problems faced by ITS Developers

Self (1985) has argued that intelligent computer-assisted learning (ICAL) was named as one of the nine research themes in the Alvey programme. Furthermore ICAL is an extremely important task area with potential mass markets and the ICAL theme is somewhat practically-oriented theme compared to the other eight. However he goes on to say "It is not clear why ICAL should be seen in this light. A handful of prototype ICAL systems do exist, but most are incomplete demonstrations and the others show that we lack the knowledge necessary to carry out successful implementations". Present ITSs claim to combine AI, psychological models of the student and the expert and educational theory. This, however, does not preclude most of the so-called Intelligent Tutoring Systems from criticisms. In writing about this field we have already specified some these drawbacks.

One major criticism is whether ITSs could be truly described as "intelligent" systems. Nwana (1990) cautions that some researchers are hesitant about making such a claim and this issue is not restricted to this field, "most researchers appear to be reasonably content with the acronym ITS. This is fine as long as everyone involved with the area understands that the usage of the word 'intelligent' is, strictly speaking, a misnomer. This does not appear to be the case, resulting in some very ambitious goals/claims, particularly in the more theoretical parts of the literature: this also appears to be a valid criticism of the entire AI literature". Clearly however, many of the characteristics outlined previously (especially separation of pedagogical and subject knowledge as well as diagnostic error rules) mean that an ITS is capable of changing its teaching behaviour based on how well students seem to learn from one strategy versus another. This certainly shows that intelligent tutoring systems can be considered as adaptive or self-modifying systems.

However, the point of this discussion is to show that while CAI can be considered a mature technology, ITSs are currently at a pre-technology phase. This is because they require a lot of labour-intensive planning in their development. Notably they require a more detailed analysis of the subject matter to be undertaken. For example, while an hour of CAI would require 100 hours of an experienced programmer's time to produce, ITS development could possibly take twice as long. A consequence of it being a relatively new field and not yet established it suffers from being both derivative and innovative. While on the one hand, Self (1988) argues that "ICAI researchers bring with them or adopt theories and methodologies from associated disciplines such as psychology and computer science. This interdisciplinary is, as usual, a strength and a problem....we have different, sometimes conflicting, perspectives on the same issue". On the other hand, "ICAI is innovative in that....it contributes back to associated disciplines and also - as it must if it is to justify its own label - generates research questions of its own".

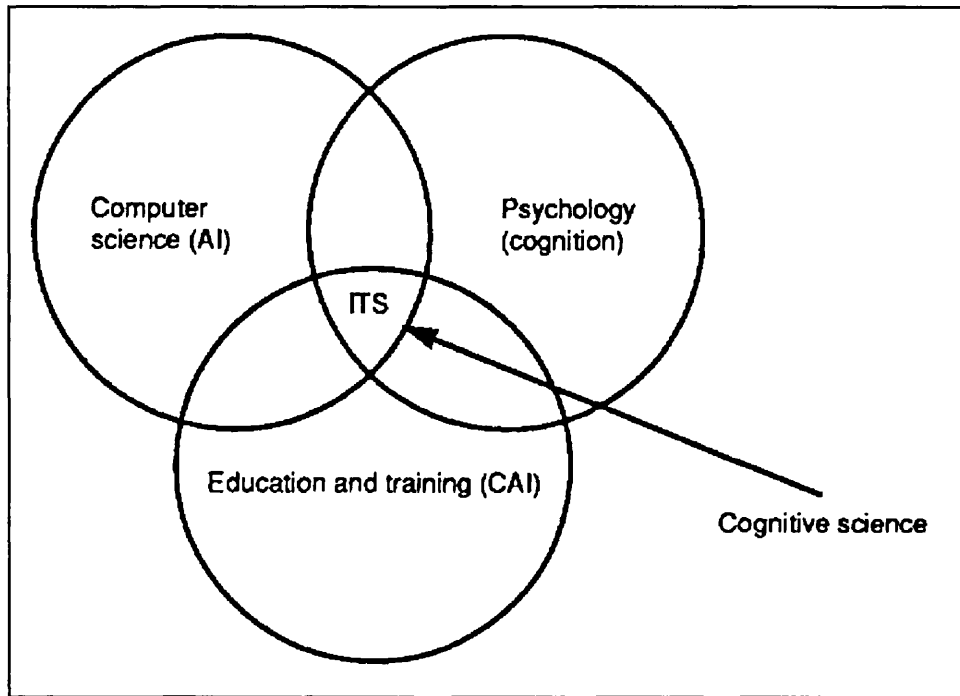


Figure 3.4: ITS Domains (from Nwana, 1990)

In addition, building an AI-based system from scratch requires a lot of specialist expertise. As Nwana (1990) notes "The design and development of such tutors lie at the intersection of computer science, cognitive psychology and educational research; this intersecting area is normally referred to as cognitive science". Accordingly this field requires the expertise of at least one person from each of its subject areas, namely a computer scientist, psychologist and educationalist. Nwana goes on specify what the other significance is of ITS research spanning three different disciplines, "It means that there are major differences in research goals, terminology, theoretical frameworks, and emphases amongst ITS researchers....ITS research also requires a mutual understanding of the three disciplines involved, a very stressful demand given the problems of keeping abreast with even a single discipline today". Clearly then the implementation of any ICAI system demands considerable subject knowledge.

Like most AI programs, ITSs tend to be very computationally complex and require enormous amounts of memory. For this reason, they usually have been developed on high-performance machines. Even then, the response times of ITSs are often very slow and unacceptable for operational environments. As Ross (1987) writes "Systems such as these

call for many of the techniques that AI has made its own, such as knowledge representation, specialised searching and matching techniques, inference and dialogue handling. Unfortunately these tend to consume a great deal of computing resources, so preventing widespread economic use. ITSs tend to be very hard to construct, too, and so their practical use can only be economically justified in special cases".

Moreover this field suffers from a lack of software tools for building ITSs. As Yazdani (1986) indicates "This means that the production of an ITS would be at least an order of magnitude more expensive than the production of a CAI system". He emphasises that the ITS field suffers from another drawback, "While the development of traditional CAI systems is greatly facilitated by the use of author languages, construction of an ITS still seems to be a one-off process". This is primarily because ITS requires powerful knowledge representation formalisms but also due to the rich assortment in methods of design.

Yazdani (1986) reports that "Most ITSs (see Sleeman and Brown, 1982) have been designed and remain as prototypes. There are some notable exceptions among them: SOPHIE (Brown, Burton and deKleer, 1982) is the oldest and most well known. SOPHIE was sponsored by the US Department of Defense (AFHRL, ARPA, TriServices). After limited use for on-site job training over the ARPA network for two years it is now no longer maintained". As a consequence most systems that are published in the literature are prototypes, developed largely as PhD projects. For example, Yazdani (1986) cited only two commercially available ITSs (not surprisingly these were in the domain of computer programming) and little has changed in the intervening years. A large number of these are rarely appraised, let alone evaluated for future research. This basically means most systems suffer a vicious circle, remaining at best prototypes, which are developed and then thrown away. This is currently the greatest problem i.e. not many fully operational ITSs.

Moreover there are a number of unresolved problems in developing advanced computer tutors: how to provide an acceptable "man-machine interface" (e.g. by the use of natural language or visual displays), how to increase motivation of the users (by well-designed learning environments increasing social interaction), how to be more sympathetic to first-

time users (e.g. by a dynamic on-line help facilities which change with experience)? But these are limitations of current practice, not of principle: a few systems already exist which, to some degree, do not have these limitations. These limitations are imposed by our own incompetence and lack of imagination. As Ross (1987) suggests "the AI world is fairly inward-looking when it comes to picking experimental domains.

Most of these do a lot less than the topic indication suggests and are only experimental, with the earliest ones being in general the least sophisticated. Hardly any have been properly tested on more than a very few people. Fortunately, computing power is becoming sufficiently cheap and powerful that we can expect to see some interesting tests of the merits of ITSs within the next decade".

Although the raw power of the computers used for AI applications is an important factor in determining performance, the efficiency of the programming involved is also a consideration. As discussed earlier, past ITSs have been developed as research tools, not as operational software. Hence, relatively little attention has been given to making ITSs run efficiently. If standard techniques used in data processing were applied to ITSs (e.g. code compression, hash tables, etc.) or the programs rewritten in general-purpose languages that run faster, substantial improvement in performance could be achieved.

Knowledge of how to build an ITS is not well understood and we must attend to detail and intuitions about teaching and learning must be researched if we are to build more sophisticated machine tutors, for example those that will reason about a student's knowledge, monitor their solutions, and custom-tailor their teaching strategies to their individual learning pattern. Indeed one of the most notable researchers, Nwana (1993), pointed out that even today evaluations are not taken seriously, "Evaluation is still a relative taboo activity in AI research; it is our belief that if others were able to reconstruct the systems then more of them could be appraised. It is currently the case that most systems remain as prototypes which are rarely evaluated. Since these prototypes also seldom go out of the laboratories in which they are developed, their research value is largely lost, unless the researchers continue the work. However, if enough was published

of the system to⁶ as to be reconstructable, it is likely the research value of these systems may be preserved as *much*⁷ more other researchers may base their future research on what others have accomplished in theirs. The benefits to be accrued by reporting more of our systems can not thus be overstated." (15;18;20;5;18;18;25;13)

3.5 Why evaluate intelligent tutoring systems?

Littman and Soloway (1988) have questioned the need for evaluation. For instance, in their own work to build PROUST (an ITS which finds nonsyntactic bugs in students' PASCAL programs and then tutors the students about these errors) they found evaluation costly, frustrating, and time-consuming. The evaluation consumed nearly as much effort as the design of PROUST itself. But we suggest there are at least two main answers to why evaluation is so important.

Firstly, we can turn Littman and Soloway's argument on its head by suggesting that by investing so much time and effort in producing an ITS, surely it would be a waste not to see it working fully in an applied setting. Certainly, learning from the mistakes would improve its life-span as well as its usability. Indeed, Littman and Soloway (1988) themselves acknowledge this: "We have learned a great deal that we might not otherwise have learned about how novices learn to program, how to teach programming, and how to build ITSs to actually do the teaching." Another pair of researchers working in this area, Shute and Regian (1993), express the same enthusiasm for evaluation, "Building a tutor and not evaluating it is like building a boat and not taking it in the water. We find the evaluation as exciting as the process of developing the ITS. Often, the results are surprising, and sometimes they are humbling. With careful experimental design, they will always be informative". Clearly this is reason in itself for embarking on this venture.

⁶[sic] = system as

⁷[sic] = many

Ford (1988) relates an argument which occurs regularly within the artificial intelligence community, "The develop-test-review-throwaway sequence evident in much of AI research is largely counter-productive and unnecessary. It encourages an undisciplined approach to software development that inevitably reduces the possibility of the originator or others refining his ideas through the developed software since it may back the necessary structure and modularity that are needed, as well as being incomprehensive". Consequently the second reason for carrying out ITS evaluations is to focus the attention away from short-term delivery and towards issues of appropriateness, usability and quality in system design. Littman and Soloway (1988) have referred to this evaluation gap in the ITS literature and say its effect is likely to slow down the development of ITS, "Because building ITSs is still somewhat an art, and because there are few ITSs that can be called 'finished,' designers of ITSs are currently more concerned with usefully guiding the development of their systems than with determining whether they are effective educational end products." So the field would greatly benefit from shared experience and avoid continuously reinventing the wheel.

The current informal evaluations are useful, but to progress ITS must develop formal evaluation methods. Moreover much of ITS work to date has focused on particular aspects of an overall system, such as the problems associated with student modelling and subject representation. There is a need for such work it should not be absolved from being appraised in a wider context, e.g. whether a student model is suitable for tailoring explanations and selecting appropriate tasks, i.e. change from experimental phase of the need to satisfy requirements of speed of response, cost and so on to answering more fundamental questions concerning a systems ability to, in general terms, foster learning.

Littman et al. (1988) emphasized the need to develop a more systematic approach to the evaluation of ITS which means ensuring availability of evaluation methods and also quality control. Winne (1993) underlined this point by recently forecasting that "in a future where [AI-ED systems] may be widely available to schools and training institutes, evaluations will shape what and how people learn, and what they become able to do". So why is there a lack of proper ITS evaluation studies? There are number of reasons for this, all revolving around the fact ITS is at a pre-technology stage (i.e. relatively new field):

- (a) lack of fully functioning ITSs (namely they are not robust enough to be evaluated);
- (b) lack of accepted formal research methods;
- (c) difficult to assess - since traditional methods less appropriate for assessing minor changes;
- (d) interdisciplinary make-up of ITS community means different standards and/or lack of interest in evaluation.

3.6 How should ITSs be evaluated?

We have already indicated the sheer variety of ITSs and that their high development costs dictate that methods must be devised to evaluate them. However this is not a trivial task, as Mark and Greer (1993) pointed out: "There are few agreed upon standards within the ITS community to guide investigators who wish to evaluate systems. However, other fields have developed evaluation techniques which may be applicable to ITSs." With this in mind we have summarised a few notable views on the types of evaluation which are deemed to be appropriate.

3.6.1 Mark and Greer's (1993) Techniques for ITS Evaluation

Mark and Greer's paper provides a brief review of techniques from intelligent tutoring systems research, expert systems design, computer-based instruction, education, and psychology which could be appropriate for formative or summative ITSs evaluation. *Formative* evaluation is the examination of a system under development, to identify problems and guide modifications and *summative* evaluation is to support formal claims about the construction, behaviour of, and outcomes associated with a completed system. Their findings are summarized under the following table:-

Evaluation Techniques	Suitability for ITS Evaluation	Type of ITS Evaluation	Evaluation of System or Components
Proof of correctness	Not applicable		
Criterion-based Adequacy General criteria	Poor suitability Poor suitability	Formative Formative	System or components(e.g. domain or communications)
Expert knowledge Expert review Turing test	Limited suitability Limited suitability	Formative Formative or summative	Components (e.g. domain)
Certification	Potentially suitable	Formative or summative	Complete system
Sensitivity analysis	Potentially suitable	Formative or summative	System or components (e.g. diagnostic and teaching)
Pilot testing One-to-one Small-group Field test	Recommended Recommended Recommended	Formative Formative Formative or summative	System or components (e.g. communications)
Experimental research	Limited suitability Recommended	Formative Summative	Components System or components (e.g. diagnostic, modelling, and teaching)

Table 3.2: Suitability of evaluation techniques (from Mark and Greer, 1993)

Proof of correctness is a check whether the system fulfils the desired requirements or goals or whether there is a correspondence between its structure and behaviour and its specifications. The authors found this technique to be unsuitable in the case of artificial intelligence programs, simply because the nature of AI, which deals "with analytically intractable problems represented as incompletely specified functions". A similar technique is *criterion-based evaluation*, where the requirements and specifications are drawn up almost as a checklist which are then used to compare the systems inadequacies. Mark and Greer (1993) consider this technique to be limited in usefulness except for formative evaluation, where general characteristics are important, and also when specific criteria can be described in detail and measured, such as interface design. Of course, using this

technique simply as a checklist does restrict its suitability, however this is not to say such a procedure used to provide formal guidelines should not play a larger role.

An altogether different approach is to make use of an *expert inspection* to judge whether a program meets an explicit standard level of performance. This type of evaluation is often used in the development of knowledge-based systems and the most well-known example is the Turing test which compares human and computer behaviour. In the author's opinion such inspection is useful for formative evaluation. However, it is not applicable for the overall merits of machine tutors since we do not yet understand "what constitutes good teaching". An alternative to this approach is *certification*, which would be based on techniques currently applied in identifying competent human teachers. Such a procedure would have the benefits of "an authoritative endorsement of the correctness of a program". The authors see this as a way of obtaining feedback on their strengths and weaknesses in formative evaluation and of obtaining adequacy ratings in summative evaluation. The only difficulty would be in judging the appropriate standards or criteria for identifying effective educational programs.

In *sensitivity analysis*, an examination is made of how varying information to the component or system would result in diverse responses. Mark and Greer argue this is relevant in the case of ITSs, where individualised instruction is a primary objective. However they indicate that the real utility of this technique is likely to depend on "whether measures of sensitivity could be assigned meaning (what range of behaviors are desirable, and when should they be invoked, given a system's goals), on how precisely measures of sensitivity could be determined, and on whether they could be compared."

Out of all the techniques the one given a highest approval rating by the authors is *pilot testing*. This is because like most systems design there are undoubtedly unanticipated outcomes of using ITSs with real users. Consequently, the three types of pilot tests that are identified provide opportunities to learn about different parts of the development process. One-to-one pilot tests are useful early in development. Small group testing is carried out later on with a representative group of students once the system leaves the prototyping stage and finally field testing takes place towards the end when completion

is only a short period away. Indeed they cite two important benefits of such investigations, "Sensitivity to the possibility of the unexpected effects, positive or negative, can prevent misleading conclusions from being reached. Sensitivity to the unexpected can also result in important discoveries about teaching and learning".

Although rarely used in computer science, *experimental research* is very common in education and psychology since it enables the researcher to obtain a relationship between a set of interventions and the outcome. Consequently it is particularly suited to examining the effects of teaching. Mark and Greer point out that "A variety of different experimental designs exist, including single group designs, control group designs, and quasi-experimental designs. Each design has its own strengths and weaknesses". However their view is that experimental techniques are more appropriate for summative evaluations than formative evaluations since they are likely to provide "overall conclusions, rather than acquisition of information". The reason for this is that the experimental approach is set up to answer specific research questions in terms of hypotheses, rather than point out ways of developing the design.

3.6.2 Legree, Gillis and Orey's (1993) Procedures for Internal and External ITS Evaluation

Legree et al. (1993) use the Littman and Soloway (1988) distinction between the external and internal aspects of ITS evaluation, to argue that the *external product effectiveness* is a function of an ITS impact on students and the *assessment of the internal features* are related to the architecture of the ITS, i.e. the components and processes of the tutor. They justify dividing up procedures into the two categories on the basis that methods for the product would only be appropriate to estimate the overall effectiveness of an ITS application, whereas more detailed and interpretative analyses would be required for specific improvements to the system. Therefore the two sets of procedures have different results, as they make clear that, "Estimating the product effectiveness of ITS applications requires the comparison of the instructional effectiveness of human tutors, traditional methods, and ITS applications across extensive problem domains. Evaluating the internal

components and processes of ITS applications require procedures that assess the instructional validity and impact of those components".

3.6.2.1 Process evaluation

Legree et al. (1993) have proposed four types of procedures to analyze ITS components:-

"1.) additive experimental design comparisons have been used to determine the impact of the system's components on ITS effectiveness.

2.) the diagnostic accuracy of an ITS application can be estimated through procedures that address the quality of micro-theories used by the system.

3.) the level of agreement between subject matter experts can be estimated correlationally and used for knowledge base verification.

4.) the immediate impact of feedback/instruction on the student can be estimated through lag sequential analysis procedures."

They describe each of these procedures and provide examples of their use. For the purpose of this discourse each procedure is briefly discussed in relation to its appropriateness.

Additive designs have been proposed to assess the impact of large ITS components that are able to be experimentally modified or withheld. It is an advantageous method since it allows aspects of the ITS to be manipulated "in order to directly assess their impact and overall importance". However it suffers from a number of problems, firstly it requires large numbers of students and is judged to be not cost-effective especially if the component is minor. Other criticisms include lack of sensitivity to specific modifications, production of equivocal results, and the inability to deal with the varying effect of competing components (the latter may also mask the type of results obtained).

The authors contend that the method of *diagnostic accuracy* has gained more prominence in ITS design "because subsequent pedagogical interactions are dependent on the correct recognition and interpretation of student errors". They cite four measures for evaluating the student model from Priest and Young: percentage correct diagnoses, the cumulative hit curve, the error fit measure, and the micro-theory evaluation quotient. Legree et al. (1993) add another measure, that of incorrectly diagnosed errors and non-diagnosed errors. The percentage correct diagnoses (PCD) is criticised for not representing the power of the student model "because a system could have a different explanatory production for every possible error". An adaptation of PCD gives us the second measure, the cumulative hit curve. This has the benefit of indicating when a system is nearing completion. The micro-theory evaluation quotient relies on a relatively simple theory explaining a large number of errors. However, its weakness is that "it requires extensive assumptions concerning the nature of the system's micro-theories". Error fit is described as the percent correct diagnoses measure adjusted for false alarms. The authors stress that "Unfortunately this formula confounds the frequency of correct diagnoses with the incidence of false alarms, thus specific values are ambiguous". The last category is the calculation of the percentage of incorrectly diagnosed errors and nondiagnosed errors. They are important because the former may initiate false feedback to the student and the latter is likely to be referred to an instructor to resolve the problem.

Verifying knowledge in the tutor is accurate is rarely carried out systematically. This approach uses subject experts or manuals and correlational methods "to estimate the consistency of knowledge and beliefs across experts". It has two main advantages, firstly the reliability of the values in the knowledge base can be estimated by applying coefficient alpha and secondly the responses of individual experts can be correlated to test for systematic differences between groups of experts, i.e. the correlations can be factor analyzed to identify clusters of consistent experts. However for this method to work we have to assume ratings of the experts are independent.

A significant part of the effectiveness of a tutor is the validation of the impact of ITS supplied *feedback on student performance*. One way to measure this is to calculate "a lag sequential probability", this is the ratio of actions in a specific category to the total

number of actions that occur within a specific frame following a target action. The advantage of such a measure is that a calculation can be made for either a different level of detail in ITS feedback or even for different types of feedback.

3.6.2.2 Product evaluation

In an earlier paper Legree and Gillis (1991) examined external, broad-based ITS evaluation studies to identify promising evaluation procedures that could justify continued ITS development. They analyzed past ITS evaluation studies and specified three guidelines from this review, "First, the instructional effectiveness of ITS applications, human tutors, and traditional methods need to be compared on the basis of performance data. Second, the instructional effectiveness of only extensive ITS applications should be evaluated. Third, large groups of subjects are required to precisely estimate ITS effectiveness". Although the authors refer to these three as criteria, in effect they are general guidelines. We have provided a summarised version of the authors' reasons for advocating these guidelines.

The authors contend that including three experimental conditions: performance of a classroom control, a human tutorial and an ITS condition is critical to insuring meaningful results for a number of reasons. Firstly it is illogical to construct a system to emulate a human tutor's interactions, if human tutors cannot be demonstrated to be more effective than group instruction for a specific learning objective. Secondly it is an advantage to include both controls into ITS evaluation because it permits the power of the research design to be assessed. If human tutoring and traditional instruction is found to be no different this indicates problems with the experimental design. Thirdly a comparison between ITS and human tutor effectiveness is useful because it provides an estimate of how much pedagogically important aspects of human tutor interactions have been encoded into the ITS. Fourthly should a difference between the human tutor and group (traditional) instruction be demonstrated, then the magnitude of the difference can be used to estimate the effect size of the ITS relative to traditional instruction. Since the authors contend traditional instruction is the least expensive alternative, another reason for including a

traditional instruction condition is to show that the ITS technology is at least as effective. However should this turn out to be not the case, then "it is not logical to develop these systems for practical applications".

After reviewing previous evaluations the authors concluded that most extensive ITSs have been favourably evaluated compared with limited systems. Here the authors define an extensive system as one "that can be expected to have a large impact on performance by virtue of its wide scope, while a small impact can be anticipated for a limited system". The reason is that extensive tutors are designed for actual requirements and cover a large amount of course material.

The authors put disparity in performance down to two reasons, "First, demonstrating a small effect requires much greater evaluation resources than are needed to demonstrate a large effect. Second, an ITS must be extensive in order for an evaluation to indicate a large impact on performance. An ITS application that has a small impact on learning cannot be used to justify ITS research and development". The problem here is that limited systems support very little instruction during the evaluations, while extensive tutors have a substantial effect on course performance. "The general finding that time-on-task is associated with greater instructional effectiveness is consistent with learning theory. It is noteworthy that this relationship was also observed when an extensive system was artificially shortened,". From this they concluded that performance evaluations should focus on extensive systems.

During their review the authors also found that many evaluations suffered from a problem of small groups. Of the studies they cite in the paper, all used sample sizes that were less than 20. As they indicate, "This is problematic from the standpoint of power analyses because small group sizes result in weak ITS evaluation designs. Much larger numbers of subjects need to be included in ITS evaluation designs in order to accurately assess the impact of ITS technology". The reason that sample size is so significant is that, in general, the power of a research design is monotonically related to the sample sizes used in the study and the effect size of the experimental conditions. They also discovered that the favourable evaluations utilised larger samples. Consequently they assert "These

considerations underscore the importance of evaluating extensive ITS applications with large sample sizes. For the purposes of power analysis, a conservative expectation is to assume an effect size of .80. Then it can be estimated that sample sizes greater than 34 per group are needed to ensure that the ITS evaluation will not result in a Type II error, i.e., missing the effect, at the 90 percent probability level (Cohen, 1977)".

Legree and Gillis (1991) end their review by highlighting three considerations which should be taken into account when evaluating ITSs. They suggest that ideally performance data should be measured both in terms of power (percent of correct scores) and time savings. However these two measures are negatively correlated with individuals, that is "the faster a person responds, the less likely that response is to be correct". In previous ITS (and CAI) evaluations both types of measures have been used, however they detected the emphasis has usually been on power scales. As they indicate, this is probably due to practical difficulties, "The strategy of attempting to demonstrate both time savings and test improvement can be problematic because the effect of the ITS is split across several dimensions. Whereas by holding one dimension constant, the treatment would primarily effect the other dimension....For this reason it is usually logical to limit variation one dimension and concentrate on demonstrating differences on the other dimension".

Another point worthy of attention is that tests should estimate the overall effectiveness of the ITS. So that for example, "the outcome measures should access performance on problems specifically related to the tutor, as well as on general class exams". This is intended to provide a comprehensive patterns of results. Legree et al. cite a shortcoming of many ITS evaluations, "the tests are not adequately described. Neither reliability, validity, effect size, nor variance estimates are reported. Without these values it is difficult to assess the meaning of the evaluation".

3.6.3 Murray's (1993) Evaluation Methods for Exploratory Research

Murray argues that "many of the fundamental questions facing the ITS field do not lend themselves to investigation using control groups and statistical methods because they are

still too poorly understood, or because the nature of the questions require qualitative analysis". Therefore in his paper he puts forward a number of evaluation methods applicable to *exploratory* research areas. His view is that the benefits of exploratory research do not require statistically significant sample sizes and that unlike scientific studies this type of research provides information about how to build systems that produce the required behaviour. Murray states that there are five experimental and design methodological paradigms within this category: iterative design, participatory design, formative evaluation, qualitative evaluation and the case study method. Exploratory ITS research may be based on one or a combination of these.

He presents a number of methods for exploratory research, which are given below, each of which will be briefly examined.

EVALUATION METHODS FOR EXPLORATORY RESEARCH

Outside Assessment

- On-site expert evaluation

- Panel of experts

Wizard of Oz experiments

Existence proofs

- Description of a prototype system

- Simulating human behaviour

- Toy domains

- Inductive argument

Observation and the qualitative classification of phenomena

- Procedural task analysis

- Cognitive task analysis

Structured tasks and the quantitative classification of phenomena

- Interviews, questionnaire, and surveys

- Matching, ranking, and sorting tasks

Statistical methods

Concept maps

Performance metrics

Cognitive change

Diagnostic accuracy

Feature usage

Internal Evaluation

Knowledge level analysis

Process analysis

Ablation and substitution experiments

Comparison studies

Gold standard

Theoretical corroboration

Corroboration and duplication

Table 3.3: Evaluation methods for exploratory ITS research (from Murray, 1993).

The opinions of experts or a large number (of potential) users of a system known as *outside assessment*, can be significant if there is agreement. Murray however, argues this is not usually proof of anything. Two types of such assessment are: on-site expert evaluation and panel of experts (or "blue ribbon panel"). The former is the normal way of getting experts to observe and assess the behaviour of system, for example several teachers rating an ITS effectiveness with some students using it. While the latter relies on a select group of experts being questioned by convening them or through correspondence. Both these methods are useful in substantiating the usefulness of a particular design. *Wizard of Oz* experiments use a human to simulate the behaviour of a proposed system, so that one can test aspects of a program design before taking the effort to implement it.

Existence proofs are common in the literature when used with other methods. This method bases its conclusions on the successful implementation of a system or application of a

method, to propose a new system architecture or design method, for example in a toy domain, it is possible to show a theory or computational model works in very constrained (i.e. "blocks world") situation. Clearly this method suffers from weak generalisations without other methods and is only powerful if the researcher describes the goals and assumptions of the study, discusses design tradeoffs, and documents surprises and failures.

The intention of *observation and the qualitative classification of phenomena* is to identify classes of phenomena, patterns, and trends in the interaction of people with instructional systems. As its name suggests this data is gathered primarily through observation, anywhere from natural to contrived situations. Murray states that "The observation can be of novices (learners) or experts (teachers), and can be aimed at identifying behavioral phenomena or at inferring cognitive phenomena".

A different approach is to focus the collection of data by limiting or organizing the responses of the subject. This is collectively known as *structured tasks and the quantitative classification of phenomena*. Here quantitative data is gathered in contrived or structured situations, for example by interviews, questionnaires and surveys. The benefit of this approach is that more voluminous and precise data can be obtained by using more structured tasks. However, as Murray acknowledges this is at "the risk of forcing an unnatural setting or disallowing non-expected forms of data".

Performance metrics are methods which use quantitative results to explore individual factors or features. These are important in exploratory settings when data pools are too small to allow statistically significant conclusions. But it is worth noting that such methods of "measuring *system performance* include false positive and false negative results". One example is diagnostic accuracy where measures of the success of ITS components are assessed by recording "hit and miss rates" of students.

Murray points to a category of *internal evaluation* methods which "study the relationships between a program's architecture and its behavior, and often involve in-depth analysis of program traces and data structures". Ablation and substitution experiment is one example

of such this approach, where part of the system is removed and a more primitive version is replaced and the system's performance is observed for its affect.

Comparison studies are a general class of methods which note similarities and differences between the behaviour or design of a system with a standard or even another system. Although comparisons are important on their own, they do not add new knowledge to the field and therefore Murray proposes that comparisons "are strengthened if they include empirical information and detailed comparisons of program functionality, behavior, or educational effects". The three methods suggested to achieve such comparisons are: judging system performance against a well-known successful case (i.e. "gold standard"), arguing at a theoretical level for the generality and extendability of a new approach to other systems (i.e. theoretical corroboration) and simulating the behaviour of another system (i.e. empirical corroboration and duplication).

3.7 Overview of evaluation methods

Given the variety of evaluation methods, how do we decide which one is appropriate for our purposes? From an examination of these methods two important guidelines emerged. Both were based on possible questions which would confront an ITS evaluator: *Are you evaluating the whole system or a part of the system?* and *Is it possible to use an experimental design to conduct an evaluation?* These two main questions can be used to separate the various methods. So that a method could be differentiated from a number of others on a scale between *external evaluation* (considering the whole system) and *internal evaluation* (testing a component of the system). In addition a method could be classified along an *exploratory research* versus *experimental research*. In the case of exploratory research such methods applicable where samples are small and the area is poorly understood. Evaluation methods which are categorised as experimental are likely to involve manipulation of variables and require statistically significant groups. The following classifies the methods relative to these two dimensions. This distribution should provide a selection of methods for the appropriate evaluation.

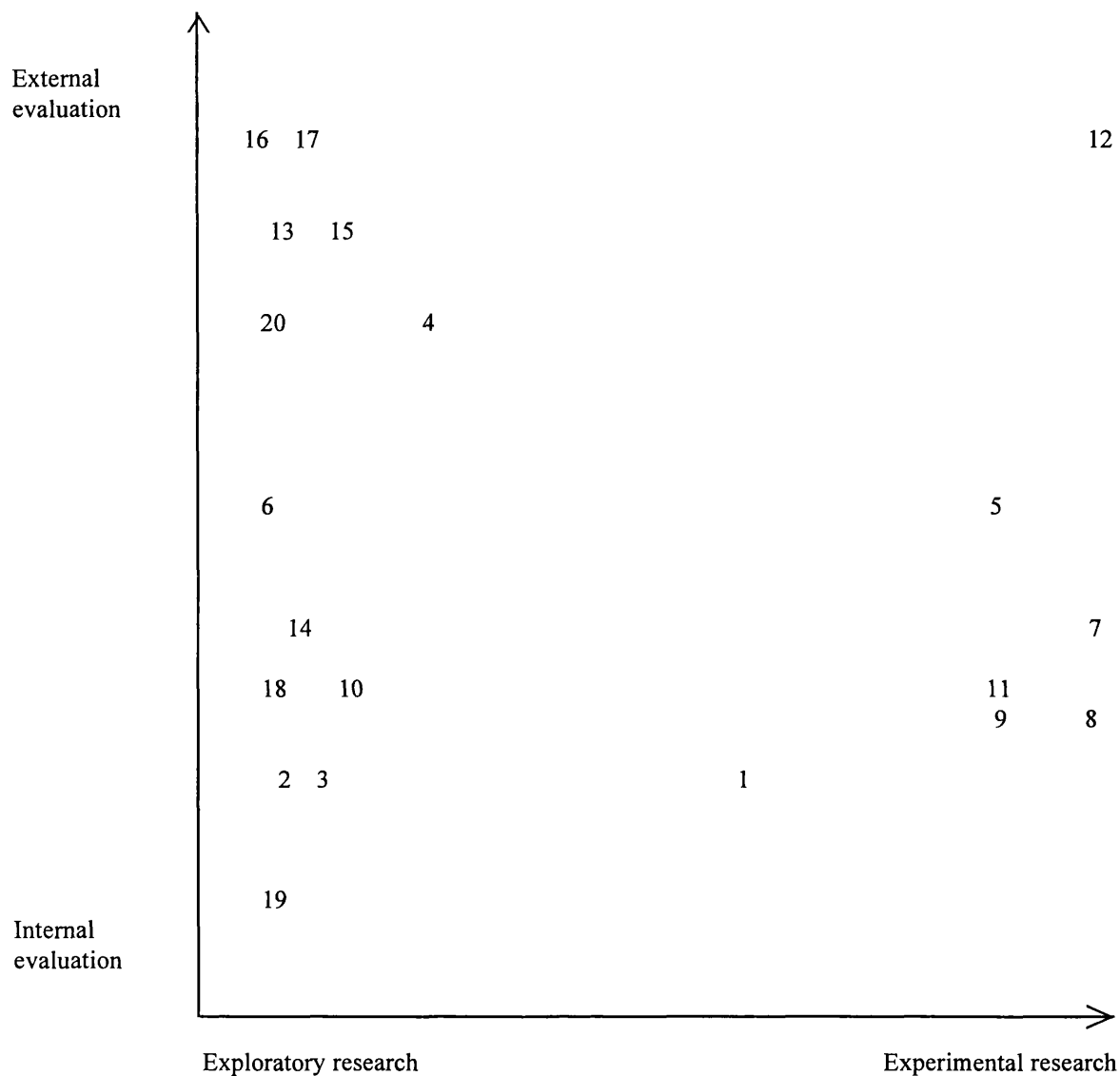


Figure 3.5: Classification chart of evaluation techniques.

KEY TO EVALUATION TECHNIQUES:

Mark and Greer

1. Proof of Correctness
2. Criterion-based
3. Expert knowledge
4. Certification
5. Sensitivity Analysis
6. Pilot testing
7. Experimental research

Legree, Gillis and Orey

8. Additive experimental design
9. Diagnostic accuracy
10. Level of agreement
11. Immediate impact
12. Product evaluation

Murray

13. Outside assessment
14. Wizard of Oz expt.
15. Existence proofs
16. Observation & qualit.
17. Structured & quantiat.
18. Performance metrics
19. Internal evaluation
20. Comparison studies

We end this discussion of ITS evaluation, by describing one of the few comprehensive evaluation models described in the literature. Simon Bento (1990) argued that there are a number of possible objectives when evaluating an ITS, i.e. assessment of capabilities, its impact, its pedagogical efficacy and most notably determining its true nature and purpose. Consequently he proposed a holistic evaluation model. It is an applied evaluation model that includes a number of integral components, all of which need to be tested. The model neatly breaks down into an *internal* evaluation sector and an *external* evaluation sector. The model also includes an additional factor which doesn't fall into either of the two sectors, namely cost-effectiveness.

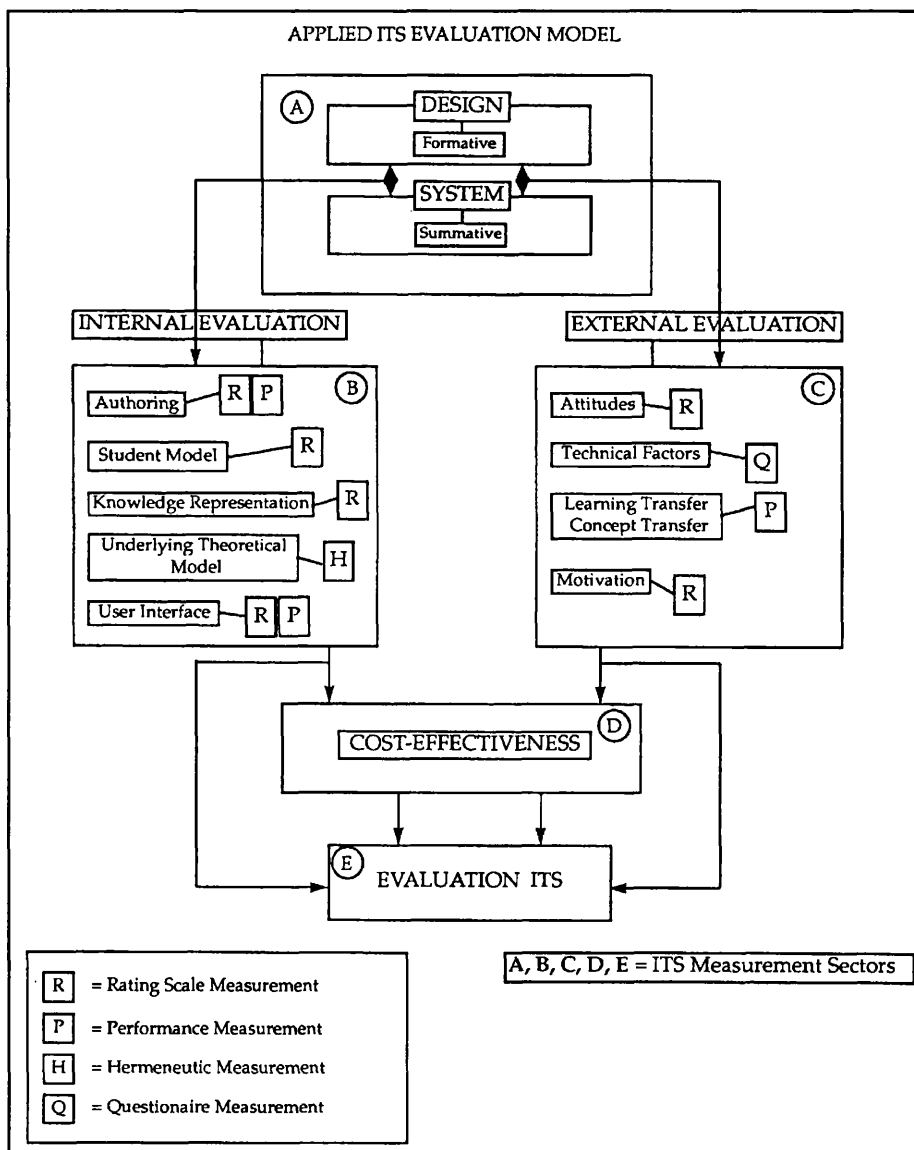


Figure 3.6: An applied holistic model for the evaluation of ITSS (from Bento, 1990).

3.8 Key Points

- Intelligent tutoring systems (ITSs) are the result of artificial intelligence (AI) techniques being applied to educational systems. They are developed as research vehicles, not as operational systems. ITSs tend to be very hard to construct so practical use is only economically justified in special cases.
- Most ITSs are largely constructed as research projects, a large number of these are rarely appraised let alone evaluated for future research. Evaluation is still a relative taboo activity in artificial intelligence. Reasons for the rarity of evaluation studies are: (i) lack of fully functioning ITSs, (ii) shortage of accepted formal research methods, (iii) difficulties in assessing minor changes in ITS, (iv) interdisciplinary make-up of ITS community (differing evaluation standards).
- Evaluation is crucial because we need: (i) to make use of the huge investment in time and effort that has been expended; and (ii) to resist the develop-test-review-throwaway culture evident in AI (i.e. improve upon them and support a longer life-span). Current informal evaluations are useful but for ITS advancement we must develop formal evaluation methods.
- We proposed a classification chart (using evaluation techniques from the literature) to help other evaluators to select the appropriate technique for their study and outlined Bento's (1990) applied holistic model, is one of the few comprehensive evaluation models in the literature.

CHAPTER FOUR

DESIGNING AND EVALUATING INTELLIGENT SYSTEMS

"If we reverse our steps and move downward in the hierarchy, decision-making is taken over by semi-automatic, then by fully automatic, routines, and with each shift of control to lower levels the subjective experience of freedom diminishes, accompanied by a dimming of awareness. Habit is the enemy of freedom; the mechanisation of habits tends towards the 'rigor mortis' of the robot-like pedant (Chapter VIII). Machines cannot become like men, but men can become like machines." [Koestler, 1989]
(1;25;5;19;1;14;4;4)

4.1 Intelligent minds or intelligent machines?

Intelligent tutoring systems draw on the techniques and expertise from the field of artificial intelligence (AI). However as we pointed out the use of the word *intelligent* has been much exaggerated. In this chapter we will examine what claims are made for artificial intelligence and how this affects both the creation and evaluation of such systems. The goal of AI is treated with much ambivalence within the domain. While some argue that the purpose of AI is to make computers do things which humans minds can do, others assert it is simply to develop more powerful computer systems. This divide is known as the *strong* and *weak* schools of AI respectively. The strong AI group share the belief that machines might one day actually think or would only behave as if they were thinking. However the weak AI group are much more interested in the construction of programs that behave in an intelligent fashion, regardless of whether or not the principles the programs employed bare any resemblance to the way a human brain would do things.

Currently, there is not enough known about what constitutes intelligence. One of the reasons for this is our lack of understanding of even human intelligence. We are just discovering how the brain and nervous system works. The other main reason is our lack of knowledge about intelligent activity, though we know it is more than just recognising patterns, making reasoned arguments and understanding language. So that, for instance,

how will we be able to tell if a computer system is *actually* intelligent, rather than merely *simulating* intelligence, and do we need to know if there is a difference? Neisser's (1963) classic paper takes up some of these issues when comparing the ways in which humans and machines *think*. If the question is phrased in a truly a narrow way, that is "Can a machine function as if it had human powers of reasoning?" then we contend that there are three possible positions we could take: the *sceptical one*, which would be "No, it's clearly impossible", the *optimist one*, that is "Yes, if you put enough rules into a computer, it will be indistinguishable from a person" and those who believe in *simple imitation*, "Yes, but only by avoiding specific rules and instead having a learning capacity which will produce a machine which is equivalent to human ability". Collins (1992) considers the possibility of actually producing a smart machine and questions what an appropriate test of intelligence would be. One of the tests he discusses is comparing the behaviour of an intelligent computer with the actions of a person, when both work through identical terminals. This was proposed by Alan Turing, a mathematician often associated AI. Suffice it to say these issues are still being hotly debated in the literature and a good introduction to these and other related issues is provided by Brown (1989) and Hays, Ford and Adams-Webber (1992).

4.2 AI hardware, tools and techniques

Although the debate about which approach is the correct one persists, the machines that may make it possible continue to evolve in terms of hardware. The *first generation* of computers were valve based machines, *second generation* computers used transistors, *third generation* computers used integrated circuits and *fourth generation* computers use large-scale integration circuits. With successive generations there have been significant improvements in reliability and processing power coupled with reductions in size, and consequently there is a belief that ensuing generations will lead to *intelligent* machines. *Fifth generation* computers are being developed which have enormously improved (parallel) hardware and will have greater reasoning powers. It is thought that by the time *sixth generation* machines appear they will possess self-adaptive learning abilities like the brain. While computers were originally developed as calculating machines, i.e. mainly as

"number crunchers" and mainstream computing has become data processing, there has been growing research interest in using computers for *thinking* (i.e. problem solving involving reasoning with concepts rather than numbers). This has come about not just by relying on more powerful machines, but through re-focusing artificial intelligence research on improving the software, particularly by employing its own variety of methods, techniques and tools.

AI research distinguishes itself from conventional computer science by the use of declarative programming languages, two of the most well known are Prolog and LISP. They specify *what* should be solved rather than *how* it is to be solved. In addition, AI languages are designed for symbolic processing, i.e. not just for manipulation of numbers but also for general abstract quantities. Similarly, the way knowledge is represented in AI systems is an important AI technique. Five basic data structures have been developed for knowledge representation: production rules, frames, semantic networks, objects and declarative logic. These methods of representation allow large quantities of data and the relationships between them, to be stored entirely separate from any reasoning (or inferencing) mechanism. Moreover AI differentiates itself by the use of a generic tool called a *shell*, which is often used to deploy problem solving strategies and to control the execution of a system. It typically consists of an inference engine (i.e. program which carries out inference, search and matching operations), together with routines for inputting rules, outputting results and providing explanations. A long-term aim of AI researchers has been that humans should be capable of interacting with a computer system in exactly the same way as another human (i.e. having a computer with the ability of learning, understanding and using natural languages such as English, Dutch or Chinese). Although there is considerable debate about whether a system capable of understanding the most subtle nuances of natural language could ever be realised, (in fact even new users required to know a lot of information even to use computers) significant progress has been made in certain specialised areas where the natural language is constrained and where technical terms form a significant proportion of the interaction. Thus natural language interfaces (or *intelligent* front-ends) have been developed for computer packages, which include natural language understanding and generation capabilities to enable a user to read and understand feedback from the computer and also to give commands to the computer easily.

A large proportion of artificial intelligence research is directed to the development and implementation of intelligent systems, for example expert systems, decision support systems and intelligent tutoring systems. These systems rely heavily on many of the methods, techniques and tools specific to the AI field (such as those mentioned previously), as well as those of storing vast quantities of knowledge in a database. The first of such systems to be developed were expert systems and are scientifically based. They are computer programs designed to emulate the decision-making abilities of one or more experts in a well-defined problem domain in a similar way to that of a human expert if one were to be consulted, such as a lawyer or accountant. More recently, they have become known as knowledge-based systems (KBSs) denoting their wider application in the commercial world to give reasoned advice and solve problems on a particular topic, and draw on a large store of knowledge about the particular domain, for example geological exploration and advising on bank loans. A KBS contains at least two components: (1) the *knowledge-base*, in which large the domain-specific knowledge is stored in the form of facts and rules, and (2) the *inference engine*, which operates on the knowledge-base, and performs logical inferences and deduces new knowledge by applying rules to facts until the posed problem is solved. There may also be other subsystems for knowledge acquisition, as well as a *special user interface*. In addition, many KBSs have an *explanation component* which can inform the user (at any point of the solution) which hypotheses they are pursuing, why they chose a particular strategy, what conclusions were drawn so far, and why these conclusions were drawn.

The discipline of constructing a knowledge-based system is sometimes known as knowledge engineering and involves "picking the brains" of a number of human experts, for example a surgeon or social security administrator. If the method of knowledge elicitation is an interview, then questions are asked to force them to spell out the algorithms behind their skills, hunches and brainstorm. Subsequently the knowledge base is built up of all the facts known to the experts, as well as any "rules of thumb" (or *heuristics*) used in the interpretation of these facts. Usually these rules are incomplete or contradictory, and various measures are adopted to resolve these inconsistencies. The advantages of KBSs are obvious, they can significantly fulfil the great demand for acquiring the expert's know-how. They can also tackle complex tasks which cannot be

solved by human experts alone and they offer an opportunity to increase our understanding about the workings of the mind through a window on an expert's reasoning processes.

The disadvantages are that KBSs require substantial amounts of development time and computing speed and power. One of the reasons for this is that the problem areas tackled are large, ill-structured and difficult to formalize in one step, therefore they are less amenable to conventional software engineering. Another reason is that the realization of KBSs are dependant on incremental advances in AI such as knowledge engineering and programming methodologies. The development of KBSs also requires the commercial availability of expert system shells to make economic sense. In addition, KBSs require a specialist environment, which provide a wide range of facilities for the building, testing and debugging of knowledge-based systems and a choice of approaches. Moreover, once a system is built it is difficult to maintain since the knowledge can become quickly out of date and is less likely to be represented as simple code.

Many may consider there is a relatively small difference between using conventional computer systems and large KBSs. While to some extent this is true, it can be argued the difference between artificial intelligence and mere computers is more about what we *do* with them. As Edwards (1991) states "both the nature of the systems and the way in which they fit into the operations of the organisation are likely to differ substantially". As a consequence it is not the computer, but the *whole system* (including the programs which tells it what to do) which we identify as *intelligent* or *apparently* intelligent.

Edwards (1991) has identified a number of differences between data processing (DP) applications and knowledge-based systems. Firstly, unlike the KBS field, the technology in typical data processing is well established enough to be certain of building a DP system for the application. Moreover, the involvement of non-computing technical specialists and high-level managers is less likely in the actual building of a DP system, whereas close involvement of one or more experts in KBS development is almost inevitable. Also, the user interface is more sophisticated in KBSs because their tasks are inherently more complex than DP systems. It is extremely difficult, if not impossible, to specify the performance objectives of a KBS in the same way as a DP system, causing implications

in terms of the need for prototyping and estimating resources for development. There is also the difficulty of replacing a KBS compared with a traditional data processing application. Finally KBSs typically contain judgements and rules of thumb, which are not correct 100% of the time, and may generate their own problems with users, their application and with system testing.

4.3 Designing artificial intelligence

As we have seen applied AI and KBSs represent a promising approach for treating difficult, ill-structured problems areas whose solution requires domain knowledge, significant expertise, much experience and heuristic rules. However KBS designers face an uphill task when developing such systems: creating a system which enables users to deal with the problem using their own paradigm, simply because problem-solving is highly individual to the selected area; attempting to represent common sense knowledge; utilization of *deep* knowledge to reason from first principles; and the difficulty of understanding of why certain decisions are obtained (or even understanding the reasoning process). Despite these difficulties we still need to know what steps are needed to make progress in this domain.

Schank (1991) provides a good overview of the issues researchers faced with when carrying out work in this area by highlighting four viewpoints of what artificial intelligence is. He begins by signifying that he is not concerned that AI needs a definition but the fact that it requires substance, (hence the title of his paper *Where's the AI?*). As a consequence Schank states that carrying out an AI activity can mean one of four possible things: (1) AI means magic bullets, (2) AI means inference engines, (3) AI means making a machine do amazing things (i.e. *gee whiz*), (4) AI means having a machine learn. The first view presumes that making a machine more efficient will somehow result in intelligence as a by-product. The second view assumes that discovering some mysterious item or technique, and dropping it into a machine, will automatically make it intelligent. While the *gee whiz* view maintains that when you are able to make a machine perform some task that it hasn't ever done before, it must be AI. The fourth view is a

long-term view and one supported by Schank. It stems from the definition of intelligence as getting better over time. Therefore the smart machine is the one that *changes* as a result of its experience.

He illustrates this confusion over whether or not you are doing *real AI* by referring to his own experience of developing AI programs. In 1976 SAM, a program which summarized, paraphrased, translated, and answered questions about newspaper stories, was built. It qualified as AI because no one had ever seen anything like it, it was solely based on the *gee whiz* criterion. However when SAM was improved (now known as FRUMP) it was not seen as AI. Despite making it faster and slightly increasing its robustness, it was simply seen as a piece of software. Furthermore, during the early 1980s when huge interest was shown in expert systems by business. Schank used the work from SAM and FRUMP to create ATRANS, a program that read international money transfer messages. He realised that this was *not* doing AI at all. This was because, all that he and the other AI people at that time were doing was essentially building commercial products (in this case, programs that someone wanted). In particular they were being asked to build a shell, namely an environment for developing software.

AI has always been concerned with the solution of the deep problem behind the small problem. However, prior to 1982 AI people worked on programs using only a few examples or *limited* the domain, which resulted in the problem of scale-up. The problem lay in the question of whether writing a program to solve a specific problem would ever help in solving problems in general. Schank argues that AI programs should scale-up. In addition they should contain AI ideas, i.e. making them work should tell us something about human behaviour. Furthermore, an AI program should not just be intended to accomplish a particular task but rather to help shed light on solutions for a set of tasks. Schank's involvement in building expert systems made him realise AI must be in the size. AI must entail massive software engineering. Paraphrasing Thomas Edison he says "AI is 1-percent inspiration and 99-percent perspiration". In other words if you want to do *real AI* you must be prepared to do a great deal of work that doesn't look like AI. It also made him understand that the AI wasn't in the inference engine (which was only simple pieces of software testing the logic of rules). An expert's reasoning ability at any moment can

be modelled by a static system, however experts don't remain experts for long without learning something from experience and changing in some ways. Therefore, in complex systems AI is present in the organization and representation of knowledge, the attempt to understand the domain under study, and the crystallization of what was important in the domain and how experts in the domain reasoned. He also argues AI was in accumulating the actual experiences of the experts and indexing them.

Schank recognised that a number of issues arise when AI tries to scale-up. Scale-up is not simply a matter of increasing the order of magnitude. It also includes dealing with real problems. This, in turn, means increasing the complexity since we have to account for all the myriad of possibilities provided by reality. However, he is ready to acknowledge that solely because a system is small does not need necessarily mean that it lacks AI. Since AI is about the representation of knowledge, even a small functioning computer program that someone actually wanted, could be an AI program, if it were *based* on AI ideas. If issues of knowledge representation were addressed in some coherent fashion within a given program, AI people could claim that it was an AI program. He even goes so far as to say that, if AI people don't deal with these issues there will never be any AI. He presents five practical problems that must be faced before carrying out any real (scaled-up) AI:-

- (a) Real problems are needed for prototyping. It is not possible to continue working in toy domains.
- (b) Real knowledge that real domain experts have (including contradictory viewpoints, exceptions and intuitive feelings of the problem) must be found and stored.
- (c) Software engineering is harder than you think.
- (d) The focus must switch from simply producing researchers to producing more builders.
- (e) Getting the field to the next level of theory means passing on the skills of how to build AI systems. So that knowledge is more available, to make the next system in the domain that much easier to build.

In summary, where is the AI? Schank indicates AI is there in the *size*, the *ideas*, and the *understanding of what contributes* to the behaviour of intelligent beings.

In fact it can be argued that AI does not quite fit in with other categories of science. Although AI is often associated with cognitive psychology and can be compared with mathematical and engineering disciplines, Campbell (1986) argues its nature simply does not conform with any of them. The reasons he gives for this are that, unlike other sciences, AI refers to objects created by humans and unlike most technologies AI does not have its parameters set by the sciences of the materials which they are created. Even parallels, of AI, with pure mathematics are found wanting. He concludes that AI can be regarded as a protoscience, with a heavy reliance on the application of heuristics. This has implications for the testing of theories in AI. He contends that testing in applied AI can simply involve the running of a program which embodies a model. If this model produces results which *parallel* those of human behaviour then the model can be said to be successfully proved. However this idea can be fraught with difficulties since a model may carry out an *intelligent* job (matching the performance of a human) but have nothing in common with human methods. In the following section we shall discuss some of these kinds of problems.

4.4 Evaluating intelligent systems

Are AI systems so different to evaluate than other systems? This section attempts to answer this question. Certainly, the fact that developers are seeking the appropriate methods and techniques to evaluate their AI systems, in an area which is not well understood, suggests a cause for concern. As Cohen and Howe (1989) report "Evaluation is not standard practice in part because we do not have formal research methods, standard experiment designs, and analytic tools. Where will they come from? The fundamental challenge is that we must develop them ourselves. We must refine existing rudimentary evaluation practices and develop new ones". It also must be emphasised that these research methods must be relevant enough to tackle the questions in this field. Cohen and Howe (1989) describe a problem specific to this area, "since AI systems are unique artifacts, and

we rarely run statistical experiments on them, how general can our empirical conclusions be? When one tests hypotheses by comparing, say, groups of undergraduates, one can be reasonably sure the results will hold for the population at large. Can we ever be convinced that the results of an experiment on an individual AI system are general to other systems? The criterion for accepting a result in statistical hypothesis testing is that it almost certainly did not occur by chance. Do we have a criterion as convincing as this in AI?"

In the following pages we will present three different perspectives taken to evaluate intelligent learning environments each from a different paper, followed by a discussion of some of the salient issues involved.

4.4.1 How to carry out a "controlled" evaluation

Shute and Regian (1993) describe how few intelligent tutoring systems in the literature have been subjected to rigorous, controlled evaluations. The results from those that have been evaluated in this manner demonstrate these tutors do accelerate learning with no degradation in outcome performance compared to appropriate control groups. However, they are not only concerned with the lack of interest in evaluation but the fact that only studies indicating successful results are being published. Shute and Regian put this down to the interdisciplinary makeup of the ITS community and a requirement for appropriate training. In order to provide some guidance, they maintain what is required is "to evaluate the efficacy of an ITS. Efficacy, in this paper, refers to assessing if the system teaches what it was intended to teach, to what degree, in comparison to what, and at what cost". Primarily the authors provide seven principles which are believed to underlie a good ITS evaluation study.

Before outlining the seven principles we need to review the general approach to research and development. The aim of experimental design is to the measure of effect produced by some independent variable on a dependent variable. Shute and Regian point out that "If the casual link between independent manipulations and dependent measures is equivocal, the experiment is said to lack internal validity. If the ability to generalize from the

experimental sample to the population of interest is equivocal, the experiment is said to lack external validity". The authors argue that internal validity is easier to achieve in an evaluation conducted within "a controlled, laboratory setting (addressing more basic research questions)", whereas external validity is easier to attain in an "evaluation experiment conducted in 'the field' or natural setting (more applied research questions)". As we have indicated earlier this represents the classic tradeoff between internal and external validity. So that while research in field settings (e.g., high school classrooms) is preferred since all aspects of the target situation exist, so do the potential confounds. Consequently research in laboratory settings is often selected because of the extreme control that is possible. However results in the laboratory are open to accusations of artificiality and lack of generalization. The authors believe that neither laboratory nor field *research* alone will give a complete and accurate picture of the instructional effectiveness of a particular intervention. Indeed they regard "research on pedagogy should be driven by theory and constrained by empirical observation". Here theory refers a coherent, plausible body of ideas about how people acquire, store, retrieve and apply knowledge and skill. This balanced view of the relationship between theory and data colours the approach taken by the authors to evaluation.

The authors argue that in their experience evaluation studies fail not because the efficacy of the ITS, but rather due to poor experimental design, inadequately operationalized constructs and measures, or deficient logistical planning and implementation. Therefore they present seven main principles that may be used to design, plan, and implement an effective ITS evaluation:-

- (1) Delineate the goals of the tutor,
- (2) Define the goals of the evaluation study,
- (3) Select the appropriate design to meet the defined goals,
- (4) Instantiate the design with appropriate measures, number and type of subjects, and control conditions,
- (5) Make careful logistical preparations for conducting the study,
- (6) Pilot test the tutor and the study, and
- (7) Determine the primary data analyses as you plan the study.

The first principle becomes apparent when the instructional goals have shifted (subtly or markedly) over the developmental life-cycle of the ITS. Therefore it is a good idea to review these goals and ensure the evaluation designer is intimately knowledgeable about the following critical issues:-

What instructional approach underlies the tutor? - Is it a system with pedagogical intelligence, a coached-practice environment, or a more free-form discovery microworld? Is the system supposed to guide learning, or provide a rich environment for the induction of principles, or allow students to practice skills?

What learning theory does it assume? - Is there a clear knowledge or skill-acquisition theory in the literature that motivates the instructional approach of the tutor?

What exactly does it teach? - Specific and measurable knowledge or skills should be expressed clearly as the desired learning outcomes.

What other impacts is it expected to have? - What other ways is the tutor expected to affect the student? For example, influencing perceived self-efficacy.

In what context is it supposed to operate? - Is the system intended to supplement a lecture or laboratory, or provide stand-alone instruction? Does the system teach to individuals or small groups? What prior knowledge is assumed of the students?

The second principle means thinking carefully about the goals of the study. Selection of an appropriate experimental design will depend on knowing what you want to find out from the research. Some of the following questions are worth considering:-

What would you like to know after the study is completed? - Do you want to know if the ITS improves the students' present knowledge and skills, or whether the tutor affects the students' capabilities to perform some other task, or even if the tutor is better than conventional classroom instruction using the same material (i.e. a comparative study).

By what standards will you measure success? - You need to identify ways to measure whatever is being taught (e.g., indices to assess the veracity of knowledge, or successful application of problem-solving skills) and to whom your students will be compared on these measures.

What are potential confounds, and which of these can you control? - Pinpointing potential confounds before conducting the study makes it easier to control them (beforehand, by altering the design, or afterwards, statistically).

Will you use quantitative indices, protocols, or observational data? - These three types of data represent the most prevalent means of capturing what a student is learning from the ITS.

The third principle involves choosing an appropriate design to test your research questions. Are you conducting a *formative* or a *summative* evaluation of your system? While formative evaluations have an internal control condition and suggest how can the system be made better, summative evaluations have an external control condition and can indicate how this system compares to some other system or approaches. Formative evaluations are conducted during the developmental phase of an ITS to find weaknesses early enough so that design changes can be implemented. For instance the authors conducted a pilot study with *real* subjects and discovered that many of them had significant difficulties learning the programming curriculum because they lacked (or forgot) some prerequisite knowledge presumed by the system (e.g., not knowing what an integer or a variable was). So they built a *pre tutor*, an approximately 2-hour computer-assisted instruction (CAI) module that instructed those 10 concepts.

Summative evaluations take place at the end of ITS development, or at the end of major development stages, and assess various aspects of the finished product. The authors present five different designs that are suitable for summative evaluation studies (some of these will be covered in chapter seven):-

- (1) Within-system Design - How do two or more alternative versions of a single tutor compare to one another?
- (2) Between-system Design - How effective is your tutor in relation to another one teaching the same subject matter?
- (3) Benchmark Design - How does your tutor fare in relation to some standard instructional approach?
- (4) Hybrid Design - A combination of the above options, and
- (5) Quasi-experimental Design - How well does your system operate in a real-world setting?

The fourth stage is to carefully plan the details of the design by considering and instantiating the dependent (outcome) and independent (manipulated) measures, the number and type of subjects needed in the experiment, and the appropriate control group(s). Shute et al. argue that it is very important to judge which criterion tasks (or other dependent measures assessing knowledge and skill acquisition) are needed in your evaluation. The dependent measures are required to reflect the goals of the ITS and related to the goals of the study. Moreover they recommend you should use multiple dependent measures, "Firstly, because ITS instruction is done on computers, you have the option to capture as much data, of whatever kind, you choose. You should err on the side of gathering too much data....Second, it is in the nature of learning and instructional research that the effectiveness of an intervention will depend, in part, on the aspects of performance you are trying to teach, and how you measure these indicators of performance". The authors provide examples of the kinds of learning or performance measures that can be collected, i.e. performance latency, performance accuracy, declarative knowledge, procedural knowledge, and procedural skill. They also caution that "Teaching one thing and measuring another is bound to result in a failed study".

With any learning task, differences between individuals are significant. Hence individual difference measures must be collected, such as profiles of knowledge, skills and traits. Several reasons are given by the authors for collecting individual differences measures in an experimental study "First, you can be sure that your treatment effects are real, and not simply an artifact of differential learner traits. Second, if you have collected aptitude data and find that they do impact the treatment condition, if necessary, you may then statistically control for those data that affect the treatment condition. Finally, if you don't have any aptitude data, then you cannot investigate aptitude-treatment interactions". Ensuring that the evaluation study is *rigorously controlled* means overcoming one of the biggest obstacles, that of identifying suitable control conditions. Since various uncontrolled conditions and unanticipated interactions appear across settings (e.g. different instructional materials, classroom dynamics, and teachers' personalities) the treatment effects can be masked. Therefore the authors offer a number of guidelines:-

- (a) Use tutors that are based on a theoretically principled approach to learning and instruction;
- (b) Give preference to data collected at a single site (rather than multiple sites) with standard procedures and measures;
- (c) Obtain a range of demographic and aptitude measures from subjects; and
- (d) Pre-specify a standard criterion task along with multiple dependent measures to be taken at various intervals during the course of learning.

Shute et al. emphasise that acquiring the right *type* and *number* of subjects (i.e. participants in research) is paramount for an ITS evaluation and with this in mind they make a number of recommendations:-

- (a) identify the target population for which the tutor is intended (e.g. university students taking an introductory Astronomy course) and make sure that the sample you are testing matches your target population;

(b) calculate how many subjects are needed for the study (as a general rule-of-thumb, ITS evaluations should have at least 30 subjects per condition for simple treatment comparisons);

(c) studies using individual difference measures as independent variables should use about 100 subjects per treatment, however this rule-of-thumb can be relaxed somewhat for sufficiently powerful designs involving extreme groups, or matched cases;

(d) random assignment of subject to conditions is critically important and should be achieved whenever possible.

Principle five deals with making sure the study has been planned properly. Example of poor logistical preparations, include failure of subjects to provide a critical piece of data or sufficient materials are not available at the data collection site. These kinds of calamity can ruin good studies and render expensive data useless. They can be avoided with careful thought beforehand and also by considering (in advance) the worst possible scenarios, such as what you would do in case your hardware or software fails. Many possible obstacles can be avoided by piloting the study. As Shute et al. say, "It is important to find these things out before committing to the expense and trouble of a full evaluation". The authors cite a few things to look for during the pilot test of the tutor:

(a) Is the tutor running bug-free?

(b) Do subjects know what they should be doing at all times?

(c) Are subjects learning anything?

(d) Do subjects indicate that they like the system?

(e) Did you estimate the learning time appropriately, or do subjects take longer (or less time) than you had anticipated? and

(f) Were all subjects able to complete the tutor?

Finally, principle seven requires that the primary data analysis should be planned before carrying out the study. A number of possible statistical techniques are presented and the authors clearly state that "Some kinds of statistical analyses are better suited to certain

classes of design types than others". Three categories of statistical techniques are worth noting:-

Confirmatory Data Analyses - when you have a specific hypothesis you want to test;

Exploratory Data Analyses - is an interactive and iterative type of data analysis, with no fixed procedure to analyze the data and its purpose is to leave the door open to alternative patterns that may exist in your data;

Cost-Benefit Analyses - cover the methods which estimate both cost and utility of systems.

Many of the principles outlined are, surprisingly enough, no more than common sense guidelines. The authors have had experience of a large number of unreported studies, which remain unpublished largely because they failed. They explain their main reason for proposing these seven principles was in the hope of reducing the number of flawed future ITS studies.

4.4.2 Informal evaluations - the alternative to "controlled" evaluation

Twidale (1993) emphasises that problems may arise when using rigorous experimental methods and instead expounds the usefulness of informal techniques. He maintains that rigorous, formal, experimental, summative evaluation (i.e. controlled evaluation) is often assumed to be the proper method and any alternative is seen as a poor one. However the author contends there are a number of arguments against this:-

- (1) Rigorous experiments are large, slow and costly;
- (2) a controlled experiment only really measures one thing;
- (3) a controlled experiment produces averaged out figures of overall performance;
- (4) unexpected interactions may lead to misleading results;
- (5) the effect of the interface can be overwhelming;
- (6) learning to learn with an ILE (intelligent learning environment) takes time.

Normally, experiments are associated with summative evaluation while informal techniques are most likely to be used for formative evaluation. Controlled evaluation is more frequently used because of its claim to be a higher status of scientific objectivity and reproducibility. Furthermore research projects, that outside bodies are likely to fund, are ones with controlled evaluative research because it produces objective measures of effectiveness. Consequently, experiments are chosen because they are judged to be the most desirable method. There are, however, also other factors which play a role in this decision-making.

The author suggests that three of the most common disciplines in intelligent learning environments (i.e. computing, psychology and education) represent their own research paradigms. Research in psychology and education belong to the scientific paradigm and are more likely to emphasise the formal objective summative experiment, whereas parts of computing have more affinity with the engineering paradigm (which derives proof that a system works from its construction). A number of issues result from this culture clash. The engineering paradigm necessitates making decisions between design options and involves trade-offs on a number of dimensions (i.e. intuitive decision-making). Experiments are usually only useful when restricting their scope to a few significant decisions within any single design (i.e. formal decision-making).

Nonetheless, there are several reasons why the scientific paradigm of repeatable experiments is less likely to happen in the general field of computing or specifically Intelligent Learning Environments. Firstly, experiments take a lot of time to set up and need to build up upon earlier ones to allow comparisons. Rapid improvements in hardware and software can reduce the value of the information gained. Experiments which involve computer systems can also be criticised because these systems are likely to develop increasing sophistication. Consequently, unexpected or unwanted results can be blamed on some technical flaw such as a primitive interface feature. Similarly, previous negative conclusions can be overruled on better features incorporated on a newly developed system. In addition, it is very expensive to reproduce systems developed on particular hardware and using particular combinations of software. Another reason may be that computing

researchers are more inclined to build and to improve systems, rather than running comparative experiments.

Informal techniques, on the other hand, are particularly useful in formative evaluation where the system is incomplete and so assessment of overall performance is inappropriate. In comparison to summative evaluations, the results from informal evaluations are more likely to be prescriptive, namely describing implications for future developments from the shortcomings of the present system. One example of an informal technique is the individual case-based study, which is informative during instances where the system needs to be studied in detail for small, incremental changes, especially when performance is less than ideal. A different technique is the *Wizard of Oz* method which is popularly used when experiments need to be undertaken during formative evaluation, where conditions need not be as rigorous as those in controlled evaluations. It involves the user interacting with a computer interface but the information is passed to a human *processor* in another location who does the processing and passes back the reply via the interface. This method is suitable for testing the efficacy of the interface in advance of the development of the internal components. In addition, it can be used to test individual internal components of an incomplete system. Although fewer subjects are involved in Wizard of Oz techniques, they are also more labour intensive. Twidale argues the differences between informal and controlled evaluation may be summarised by saying that one should use controlled evaluation when one wants to show the advantages of a system, and informal evaluation when one wants to reveal difficulties.

Informal techniques can appear imprecise and may be perceived as far from scientific. However, Twidale (1993) asserts there are ways to ensure greater rigour, without losing the benefits of speed and simplicity from these informal methods, "The test for the effectiveness of the evaluation technique is whether it leads to the building of better systems....A formative study should reveal problems in the system that can be corrected in subsequent versions....It should also reveal those elements that appear to be successful even in stripped down prototypical form, providing evidence that in the improved version of the system, they would be all the more effective....The acquisition of anecdotal

evidence can also be used to answer the criticism of many studies both formally experimental and not, that they are unique to their circumstances".

A significant advantage of informal techniques over controlled evaluation are that they can make unexpected discoveries. There are certain phenomena that may only be observed by sitting alongside students using the system, these are activities which can not possibly be predicted when it was being developed. An example of such a study is provided by Twidale. The EPIC system was an ILE developed in order to study learning proof construction in propositional logic. It made explicit the plans and goals of the student when problem solving by use of menus, form filling, annotation and explicit instantiation. One advantage of the in-depth study was that it revealed quite unexpected aspects of the learning process. These aspects can be quite subtle but can have a direct impact on students' interactions and learning activities with the ILE being investigated. The EPIC study revealed that certain students were able to competently perform simple exercises in a domain while still having major misconceptions about that domain, which only became apparent when they tackled harder problems (that could not be solved by their non-standard technique).

Twidale concluded that, "Although there maybe considerable pressure to evaluate an ILE using controlled experiments, researchers need to be aware of the problems that can arise with such techniques. Their frequent use as a measure of overall effectiveness can obscure significant variation in the ILE's effectiveness for individual learning events". He remarks that the field is still quite new and there still is a lot to be learned by simple observation. While there will be occasions when it will be right to use summative experiments. It is also true that in many situations it will be appropriate to use more informal techniques (e.g. in-depth studies) to supplement experiments or simply as an alternative. He explains that "The challenge is to make such studies acceptable as a way of providing meaningful evidence of the features described while retaining the technique's flexibility and low cost."

4.4.3 Cognitive evaluation

Or-Bach and Bar-On (1993) take a completely different approach. They claim that ITSs should not be evaluated from a computer system perspective, but instead from a cognitive mediation position. They agree that evaluation is vital both for a system and the advancement of this field. The crucial question, however, for them is how do we achieve this given the variety of systems, with differing goals, educational philosophy, implementation and many other relevant perspectives. Or-Bach and Bar-On (1993) demonstrate what should be evaluated and the most appropriate methodology for the evaluation. The authors' whole approach is based on their perception that the focus should be on formative and especially internal evaluation because of the current stage of ITS development. Moreover, in accordance with their constructivist stance to learning, they pay more attention to what the system actually provides to learning as opposed to measurable outcomes.

In their paper, the authors have attempted to answer three basic questions:-

- " ● Should ITS evaluation resemble the evaluation of any other computer system?
- What should be evaluated?
- What kind of methodology is the most appropriate for such evaluation?"

Or-Bach et al. (1993) answered the first question, by arguing that ITSs should be evaluated from an educational perspective and centre on internal evaluation, namely an approach which focuses on the educational goals of such systems rather than on their efficiency. The reason is that very few ITSs are functioning outside development labs, therefore neither the computer system perspective nor external evaluation is yet relevant.

Secondly, they recommended that instead of evaluating the system's functional components (such as the expert or student module), internal evaluation should concentrate on the different dimensions of the cognitive mediation process. The TALK (Tactics,

Analogy, Language and domain Knowledge) model is suggested by them as a suitable framework for evaluation since it captures the facets relevant to the learning process. TALK is a top-level framework not only suitable for designing intelligent learning environments but also for the refinement of the system modules. In this Or-Bach et al. (1993) emphasise that "The underlying assumption of a facet model is that there is no point in dealing with each facet alone, but with the product of the four facets. Each facet can be understood as a criterion for classification, while the objects of classification get a profile over the four facets". Their cognitive model consists of:-

T-Tactics - The instructional tactics facet defines the instructional tactics employed by the ITS to achieve its instructional goals.

A-Analogy - The instructional analogy facet deals with the instructional analogies provided by the environment, arranged from concrete analogies to more abstract ones.

L-Language - The interaction level can be arranged from natural language to a formal language. The alphabet can be words, numerals or pictures.

K-Knowledge types - refers to the knowledge types which are used in the instructional process. The knowledge types include: content knowledge, structural knowledge, and implementation knowledge.

The authors answered the third question by presenting a preferred methodology and illustrating how it can be employed. Their methodology is one of successive approximation and its main idea is to make a change in one of the TALK facets (i.e. change of analogy, language of interaction, type of knowledge to be conveyed, or the instructional tactics employed), and to compare the system before and after the change, on each of the facets. The evaluation is done with the goal of maximizing some property, and therefore the change can be evaluated in relation to that property. The authors describe it as an optimization technique analogous to *hill climbing*, that is the evaluator changes something, evaluates the change in comparison to the previous situation, and then decides what is the most promising direction for climbing. Or-Bach et al. (1993) used the

example of PROBIT (PROBability Intelligent Tutor) - an ITS for teaching probability-to illustrate their evaluation methodology. Their approach meant changing a specific facet (in this case the language) and employing a specific measure of success in this instance the property to be maximized was the amount of valuable information that is extracted from the learner responses.

PROBIT presents a student with probability problems where each concept is represented as a frame. The student's task is to choose the relevant concepts and to assign them values. The computer tutor uses a formal language for describing events, which enables the student to ignore less relevant details. In addition, PROBIT represents misconceptions based on expected behaviour within a certain context. The aim of the evaluation was to discover whether the language used to solve the problems was effective for the student to demonstrate their knowledge and whether it provides the system with useful information for student modelling. The methodology was implemented by modifying the system to include natural language explanations and asking the student to explain their (formal) answers and provide negative examples (i.e. those with incorrect answers). The responses in natural language eliminated some of the constraints of a formal notation or syntax. Subsequently, Or-Bach et al. (1993) describe how "The modified version was empirically tested and the learners' responses were manually analyzed and compared with PROBIT's diagnosis (based only on the formal language input)."

Their results showed some interesting and useful directions for change in student modelling and the interface. Most notably, the approach was able to reveal the close relationships between the functional components, particularly the student modelling and the learning environment. For instance Or-Bach et al. (1993) relate that "one of the students was looking for a new way of defining combinations of outcomes, that was not provided by the environment. This definition led him to a wrong answer, but revealed a misconception that couldn't be detected from his formal language answer. It is quite obvious that investigating the change induced by modifying one facet might provide some information on the uninvestigated facets as well. For example, some of the natural language explanations of the answer raised the question of how well the students understand the analogy of a 'stream' and a 'filter', which is central to PROBIT".

Moreover, the evaluation was able to provide a way to improve the evaluation itself. In this study the authors found the change in the language facet was confounded with a change in the task. Consequently, they proposed that "A better design (a more refined one) should make a distinction between the task requirements (asking the learner to explain and provide a negative example) and the language (that can be either natural or formal)". They also debated the advantages of Natural Language Processing (NLP) against the cost of its implementation in PROBIT. The authors agreed that a limited version of NLP might be a cost effective solution since natural language explanations might be too broad and ill defined. Or-Bach et al. (1993) conclude that the main benefit of using the TALK model and their methodology is that the direction for improvement is always geared to the cognitive mediation process.

4.4.4 Issues in evaluating

Each of the three perspectives (described above) examine different aspects of an intelligent system and express a different paradigm to evaluate them. Since the area is still new, there are no right or wrong answers to proceed and the choice must rest on the individual or organisation carrying out the evaluation. Winne (1993) extracted and scrutinised many of the salient issues which have afflicted the process of evaluation. His purpose was to highlight a number of these issues which arise from examining the results of an evaluation. It was also to depict the landscape of evaluating adaptive learning environment systems (ALEs). Winne opposes the influence of Ralph Tyler, an early proponent of educational evaluation, for emphasising that evaluation activities need to be measured against fixed models. Tyler chose his objectives in the context of an entire organisation, whereas Winne finds the task of evaluation is markedly different in an adaptive learning environment because "Today's ALEs are considerably less expansive, of shorter duration, less articulated with other educational factors, and probably much less forceful than the kinds of programs that Tyler and his successors had in mind as targets of today's educational evaluations". The idea of evaluation has also moved on. Indeed, Winne defines an evaluation study as "a systematic effort aimed at gathering and interpreting, in a principled way, information with which to gauge the worth or value of an instructional

enterprise". It is clear then that he lays great stress on evaluation as gauging worth or value rather than attempting to meet targets.

Winne affirms that evaluations are important in several ways, chiefly they affect the growth of future work in this field. They also carry considerable weight with sponsors and other interested bodies about the utility of such systems within the education and training field. In addition, when such systems come to be widely available in schools and training institutions, they will shape what and how people learn and what they will be able to do. In order to achieve this, Winne argued that we must strive to carry out better evaluation studies. Here he returns to his definition, when he insisted that "Viewing evaluations as invitations to learn [about people and systems] is perhaps the most profitable perception that can be nurtured". Winne's paper addressed the goal of producing a high quality evaluation by outlining five broad questions that any evaluation study must answer. The response to these topics is intended to provide a context in which the results of the evaluation maybe properly interpreted. These topics are:-

- (1) What is the system being evaluated?
- (2) What standards of worth and value are relevant in evaluating that system, and what justifies selecting those particular standards?
- (3) What data are to be gathered to describe the system?
- (4) What process(es) are used to gather that data?
- (5) What rules are used to filter data and to relate data about the system to the chosen standards of worth and value?

The first thing that needs to be established is "which system is to be evaluated". On the face of it this is a simple question. However, upon closer reflection it is not that easy since an adaptive learning environment is not a simple entity. Winne points out that an ALE consists of several components or modules and that each component has a particular function. Furthermore he explains that various levels of a system can co-exist in what is often labelled an ALE. As he points out "The ALE per se is not an educational system any more so than is a text on a library's shelf or a videotape not yet played". He breaks an ALE down into its various levels:-

Level 0 - is the ALE itself;

Level 1 - is a student working with an ALE and constitutes the most elementary unit of an instructional system;

Level 2 - includes Level 1 and other students with whom to collaborate and discuss the work and represents a collaborative learning environment;

Level 3 - builds upon the previous level and embracing exchanges between a teacher or coach and the student and constitutes human instructional support and guidance as a significant element;

Level 4 - describes the addition of classroom activities which help the student to connect skills which they have learned and also a context, so such skills may be applied to the "real" world without further instruction;

Level 5 - is signified by a student demonstrating or putting to use all that has been learnt with the ALE, peers, a teacher and other resources and is the realization of the progression of levels.

As a consequence, determining what system is being evaluated is not straightforward. Winne suggests "The evaluator should gather data, from participants as well as from the ALE's logs of data, with which to explore what the operative system is and how its various levels interact". The reason for this is that, apart from laboratory contexts (where conditions are highly controlled), it is likely that ecologically relevant evaluations will suffer from a student being exposed to some, if not all, of the different levels within the same study. This will lead to misunderstandings of the evaluation results.

At present there are only two types of worth or value that dominate the literature of ALE evaluation, namely the qualities of the ALE as an *engineered computing system* and the capability of the ALE to promote *students' achievements*. The first type relates to whether the ALE meets the design specification, for example whether the interface is

understandable to students as envisaged by the developers. Winne describes this category as "fundamentally engineering questions spiced with a bit of human preference. As such, it might be presumed that standards by which to judge such matters are relatively straightforward. This almost certainly is not the case". In Winne's view, this is because the engineering standards which could determine standards of worth and value are yet (if ever) to be agreed and also very few ALEs are available among the research community for this to be tackled. The second type of standard of worth or value concerns whether an ALE enables students to achieve goals set for them, and is considered an educational question. Even here the author is unable to see progress, firstly because "No ALE has yet grown to sufficient proportion to merit raising questions about whether the panoply of worth and values inherent in these large sets of goals might be met when students use an ALE to learn". Secondly, the author points out that, for better or worse, standards of worth and value attending achievements are multiple, heterogenous, contestable.

Winne claims that other forms of worth or value have yet to be directly addressed within this field. For instance, he distinguishes between the ability of ALEs to promote task transfer (i.e. how effective the student is able to translate from simulation to the *real* world) and learning transfer (i.e. how well the student is able to extend their knowledge and skills to the same or similar domain). He indicates there is a lack of agreement about what and how to instruct, so that task transfer and learning transfer are both optimised. He attributes this absence of a discussion beyond student achievements to the infancy of this field. So that motivational issues, learning skills and other issues in instructional psychology are barely addressed. He acknowledges that although standards for judging the qualities of the learner's behaviour are even less well agreed upon than are the standards for domain knowledge, they do have important consequences.

The author questions whether it is necessary to achieve educational goals by offering ideal engineering solutions. It may even be better if an ALE fails to reach the most favourable engineering standards or appear faulty, so that it encourages the student to go off and acquire useful knowledge in that domain from another resource such as a library. Using this argument Winne agrees with Twidale (1993), in that evaluations can benefit from

qualitative and exploratory modes of evaluation rather than focusing tightly on specific outcomes (that are usually expressed by strong experimental design).

Another issue related to future standards is the two different aspects of an ALE, that is the power of an ALE to adapt its instruction and the flexibility of an ALE to be customized. The former is how much "an ALE can be revised to accommodate a learner with less prerequisite knowledge and skills than planned for by the designer" and the latter is "the effort required to change ALE's subject matter to accommodate new information to that field". These affect the context of standards of worth and value, and relate both to engineering and student achievement in a particularly rapidly developing area. The author contends that as an ALE becomes more adaptive to a student's individual differences, so the normative standards contrasting student accomplishments become more complex. He asserts that "It seems sensible to recognize explicitly that standards of worth and value must be relative to those who pose them and those who implement the system itself". Clearly standards are in play in the various levels of an instructional system and one of the roles of an evaluator is to corroborate the basis of these in the activities and actions of people involved in the ALE-instructional system.

The type and quality of data collected from an evaluation study is vital. Winne argues that an evaluation of an ALE system should describe attributes of the system at each of its levels. The main criteria which usually guide the significance of the data obtained are their *reliability* and *validity* (Winne describes these as accuracy and appropriateness, respectively). Reliability is the extent to which data are free from error and validity is whether the data actually reflects what was intended to be measured. Winne readily admits that "Data can never entirely be free from error. Examining data's reliability means exploring what its sources of error are and what are the sizes of those errors". In the case of an ALE, the author criticises the researchers for not discussing reliability and stresses it needs to be taken into account when talking about students' individual differences and the extent that an ALE utilizes reliability when it adapts to learning differences. The most relevant aspect of validating data is construct validity and it refers to the idea that data always stands for something. Winne maintains that while conducting other types of research this can be quite problematic "because of the inherent concerns for testing a

theory", in evaluation studies the procedure of construct validation "is simpler because an evaluation is performed for clients or consumers". Quite clearly, he is making the distinction are between pure and applied research and claims that the client is responsible for selecting the standards of worth and value.

Another guideline, which Winne suggests should be used to collect data, is the "potential to test severely or to negate expectations about (a) how a system functions and (b) the kinds or levels of outcomes that the system was designed to promote". He compares this principle to Karl Popper's recommendation for testing a theory and offers two reasons why this may be significant, firstly to examine if there are undesirable side effects and secondly to scrutinize whether the system truly achieves its objectives. This can be accomplished by collecting many different types of data as well as from a myriad of sources.

Although there are researchers who have implied that there is a significant difference in methods for gathering quantitative data compared with qualitative data, Winne's view is quite adamant that these distinctions are not so clear cut and regardless of whether data are identified to be qualitative or quantitative, "an evaluation study must lead to judgements of worth or quality". Indeed he suggests it is more important to be concerned that data "be scrutinized for their capacity to convey meaning, to suggest interpretation, and to be challenged in principled ways".

The *observability* and the *marginal utility* of the data represent two practical issues that need to be tackled when determining what data needs to be assembled. The former is the effort needed to obtain the information from a complex system, such as an ALE, and the latter is whether an additional level of data adds any significant meaning to what has already been obtained. A balance must be kept between using overly sophisticated methods for collecting data, which decreases clarity about how well an ALE achieves its objectives and the relative ease of obtaining a mass of data, where the effort must go into examining whether a useful answer to a simple question can be found. The evaluator is always involved in a balancing act between the practical and the desirable. For instance, Winne argues that only by ensuring task and learning transfer has occurred can researchers claim

to "reach into the upper levels of education or training systems". However, this is not always possible to measure accurately and consequently the Winne recommends that one needs to be clear about the goals at various levels of the instructional system.

When an examination is made of what process should be used to gather data, it is commonly found that the researchers can fall into the trap of choosing between dichotomous approaches, i.e. formative versus summative evaluation, or the exploratory versus experimental methods, because it is argued that the "processes for collecting data need to articulate with the purposes for conducting the study". Winne disagrees with this since there is often a hidden implication that ALE systems should be evaluated competitively, or as he puts it as *horse race* studies. He contends that such practices have a tendency to mislead and are only useful "when they examine outcomes that are well-defined, well-accepted, and thoroughly represent the standards of worth and value of interest; and when the evaluator and the reader of the evaluation study both are clear about the level of the instructional systems that are being compared". Nevertheless he does recognize that the tension between exploratory versus experimental methods does arise from the need to obtain internal validity in contrast to the need to acquire external validity (i.e. measure the same variables in the variety and flavour of the target environment of the ALE system).

Winne submits a number of other concerns which arise from the attributes of exploratory and experimental approaches, principally that "neither approach *must* be guided by a theory, at least in the sense of 'strong' theories that are taken to be characteristic of the physical sciences". Although he argues that theory or model guided evaluations are more useful than 'What if' ad hoc approaches, it is also true that evaluators should not wear blinders, since "No model or theory is complete enough, nor is any single evaluation of an ALE thorough enough to examine all mediating and outcome variables that might be of interest". Secondly, he claims that despite the contrasting approaches there is "Nothing about a highly structured experiment precludes value in collecting data that are commonly considered qualitative, and nothing in exploratory work precludes collecting data commonly labeled quantitative". In fact obtaining many kinds of data from many sources can compensate for inherent weaknesses in individual types. Thirdly, the approach taken

should not predispose the researcher from considering a particular analysis of the data. Instead, the data itself should suggest how it ought to be examined.

Finally, the author attacks the false dichotomy pitted between formative and summative evaluation. He supports this argument by referring to the mistaken view of Scriven's original labels (recently Scriven, 1991 has put forward 10 fallacies that arose from those definitions). Winne relates there is really only one difference between the two types of evaluation, "Formative and summative evaluation do differ in function: the former serves improvement and the latter serves decision makers". But it is a fallacy that formative and summative evaluation are *intrinsically different* types of evaluation. Secondly, "the view that formative evaluation is or can properly be a much more informal process than summative evaluation" is another fallacy. If this were true then it would simply undermine the whole purpose of formative evaluations.

Although the use of statistical analysis to examine the data collected is usual in ALE evaluation studies, it can be quite problematic. This can be illustrated with reference to a classical experimental design, where a sample of students' scores is collected in comparison with the treatment of interest. The objective of the classical experimental design is to discover how much of the variance in the dependent variable (e.g. student scores) can be attributed to the arrangement or change in the independent variables (i.e. treatment). While it is assumed in the classical experimental design that all students undergoing a particular treatment in the exactly the same fashion, in ALE evaluations this is not necessarily the case since the ALE adapts tutorials to each student's individual differences during the course of a single tutorial or a series of sessions. Winne indicates that not only will tutorials will vary among students, the variance may well be amplified "if the ALE also strives to teach learning or metacognitive strategies because, as variance increases in the ways students approach learning with the ALE, so too will the forms of tutorials".

One source of this *noise* or uncontrolled variance is the individual differences among students, "for instance, variance in their general persistence at working with the ALE". Other sources of individual differences "such as background knowledge, motivation,

expertise with cognitive strategies", also contribute variance. The second element contributing to this variance come under the term, factors about which the data analyst is ignorant and include student accidents with the mouse, which the ALE interprets as genuine responses and inattention of the student because of an upcoming football match. Add to this a third component of variance, the ALE's tutoring component. This violates an assumption underlying all inferential statistics, that students are assigned *randomly* to treatment conditions (clearly not the case because the effect is correlated with each student's individual differences). This results in a problem of attributing the cause of the differences in the dependent variable. It is further complicated because it is this an interaction between the ALE and the students, which is at the heart of evaluating a Level 1 system.

The author recommends two ways to address this problem, "One is an elaboration of the general linear model, termed multilevel or hierarchical linear modeling (Raudenbush, 1988). In this statistical approach to analyzing data, bias in statistical estimates that arises due to aggregation of the kinds just discussed can be identified". According to the author, this approach remains as yet untried. The second approach is to tackle this problem by identifying where to assign variance in the way an ALE interacts with students and to model the effects of this variance on outcome variables. This usually means employing exploratory, informal, or qualitative methodologies for examining data. While such detailed information is beyond the scope of his paper, he goes on to give some general principles for examining data.

In the final note Winne poses a poignant question: if an evaluator, following a client's wishes, actually develops a profile of results of an evaluation study then "how does one reach a binary decision: purchase the ALE or don't; implement the system or wait?". He remarks that for all "the sophistication of modern methods for gathering data and techniques for analyzing data, distilling a complex array of findings into a single decision is still a difficult matter". He advises that, in addition to the clarity of goals, we need to consider how to analyse values.

The author completes his paper with a few worthwhile points. Firstly, the trade-off between control and generalizability can be a good indicator to users of what an ALE is and offers and, a view of what their local instructional system is, therefore should be visible. Secondly, the author sees the advantage of widening the exploration of standards of worth and value because he finds profit in combining the work of ALE researchers and instructional psychologists, especially in investigating the two kinds of transfer he outlined. Thirdly, he favours the disagreements and criticisms when defining the characteristic of evaluation studies. Indeed he believes such tensions are "the genesis of evaluative judgements and of scientific discovery, the precursor to potentially valuable ALE". Fourthly, he suggests that argument about qualitative versus quantitative forms of evaluation should be avoided and questions its usefulness. Instead he proposes asking "the user about the data that will make a difference, then adapting methodological tools to those purposes". He advocates that this should set evaluation on a course that will help the user, rather than seeking to advance a basic researcher's agenda and is a broad guideline for determining what data to gather in an evaluation study. Lastly, as for arguments of basic research along a dichotomous dimension, the author only believes in one true dichotomy relevant to evaluation studies: whether to adopt or reject the system.

4.5 Landscape of evaluation

We have examined three extremes of evaluation which can be broadly described as: *scientific*, *engineering* and *educational*. Shute and Regian (1993) emphasised a highly controlled experimental approach in which an evaluation should be "driven by theory and constrained by empirical research". In contrast, Twidale (1993) prefers the use of informal techniques because of the problems associated with rigorous experimental methods and the advantages of informal evaluations in dealing with incomplete systems, as well as permitting unexpected discoveries. While Or-Bach and Bar-On (1993) insist on a cognitive approach, where each facet may affect another facet and the educational goals are more important than efficiency. It is a style which is even able to improve the process of evaluation itself. Another researcher in this area presents a very similar classification. Adelman (1992) offers a multifaceted approach to evaluation of decision support systems

and expert systems, which consists of three aspects: *empirical*, *subjective* and *technical*. Empirical evaluation assesses the system's affect on performance (for example, whether users make significantly better or faster decisions or use significantly more information working with rather than without the system). Subjective evaluation refers to evaluation of the system from the perspective of potential users (for example, the user's opinions regarding system strengths and weaknesses). Technical evaluation is to establish how well the system is built and examines it from both an internal and external perspective (for example, testing the adequacy of its knowledge base).

Since this application of evaluative methods is relatively new to intelligent systems we cannot claim a preference based on usefulness. Whichever approach is used, we must be guided by Winne's (1993) useful advice. Firstly, that evaluation is to gauge worth or value rather than meeting targets. We must determine exactly what system is to be evaluated (i.e. from its parameters). The approach should attempt to address other forms of worth and value, apart from engineering standards and educational goals. Even though the evaluator is always balancing between the practical and the desirable, they ought to collect a variety of data and from numerous sources as possible. The approach must contain an element of theory or model guided evaluation. Lastly, Winne recommends the evaluation approach must be one that helps the user. The evaluation approach for the research in this thesis reflects a mixture of three different types of evaluation discussed. It is an experimental design to compare the effect of the system. Tests were given to users before and after exposure to determine changes in cognitive performance. In addition, questionnaires were designed to elicit user's responses to the system. Overall the approach taken was designed to help users judge worth or value of the system, address standards both educational and engineering and collect an assortment of data.

4.6 Key Points

- Applied artificial intelligence systems and Knowledge-based Systems (KBSs) are being developed to tackle difficult, ill-structured problem areas which require domain knowledge, significant expertise, much experience and heuristic rules. However, they have

been unable to break out from the toy domains used in laboratories, i.e. to solve “real world problems”.

- Roger Schank (1991), an AI researcher, recognised that real AI must be in the size (i.e. it must be scaled up to include field complexity) and must entail massive software engineering. If we want to develop real systems then we must be prepared to do a great deal of work that doesn’t look like traditional systems production. Since software lies at the core of many modern intelligent systems, the ability to manage its development (i.e. software engineering) is a key challenge facing all operational systems developers.

- AI systems are unique artifacts and consequently evaluation must deal with problems specific to this domain. Standard design tools, formal evaluative methods, and common technologies must become standard practice for developing all intelligent systems.

- We took three distinctive evaluation models to illustrate the possible directions that an evaluation of an intelligent learning environment can take: (i) Shute and Regian (1993), (ii) Twidale (1993) and (iii) Or-Bach and Bar-On (1993), which were broadly described as scientific, engineering and educational respectively. The evaluation approach for our research reflected a mixture of these three different types. Winne (1993) sketched out the evaluative landscape pointing out and scrutinising many of the issues raised by such evaluative models and supplied a set of useful guidelines for evaluators.

CHAPTER FIVE

EVALUATION

"The objective of education is to prepare the young to educate themselves throughout their lives". [Robert Maynard Hutchins, in "Dictionary of Quotations"] (15;3;20;15;18;2;18;9)

5.1 What is Evaluation?

Evaluation is a term which has come to mean different things to different people. In everyday usage, the term has been variously been ascribed as the determination of value, worth, or as an opinion, judgement, conclusion, selection, assessment, estimate, appraisal, critique or even review. Indeed the sheer variety of purposes for which evaluation has been used was confirmed by Bawden (1990) when he argued that three general ideas can be ascertained from the many meanings: *quantitative* ('measurement', 'valuation'), *qualitative* ('review', 'criticism'), and (perhaps stretching a point) *semi-quantitative* ('appraisal'). Unsurprisingly he found that evaluation is a complex and multifaceted set of concepts. As we shall see, evaluation in its simplest form is a test and at its most complex is a verdict and for the majority of the time the meaning spans this continuum appearing as an investigation or as a thoughtful inspection.

Even in the literature, evaluation can be confusing partly as a result of a misunderstanding with this term. But, also because the meaning of evaluation is solely defined by the beholder and their situation. For instance, in *computer science* it can mean anything from proof of correctness (by way of validation and verification of the system), expert review, or even a survey. However, for *engineers* the evaluation process can take one of two main forms. In a narrow sense evaluation is often taken to mean testing, where a component of a system is checked to see if it works as it should. Another aspect is the focus on the purely physical performance of system, for example rating of computer power in millions of instructions per second. Whereas, *managers* often take a more high-level approach and

are more concerned with the effects of the system as a whole. Kemp (1988) sums up their definition of evaluation as, "the process of deciding the worth of things and their costs in relation to that worth." Angell et al. (1991) suggests this originates from organisations having to decide whether one particular system is worth the investment or even discovering the relative merits between a number of alternative systems. They argue that judging the value of a system in this way "can be thought of in monetary terms (both benefits and costs), or in terms of usefulness or success (as opposed to failure)". Another descriptive way of using evaluation is to find out what a system *cannot achieve*, compared with what it *can accomplish*.

A researcher in this field, Winne (1993) acknowledges the view of what constitutes an evaluation study is mixed. However, Winne does go on to say that evaluation studies should be viewed as invitations to learn about the system. Rogers et al. (1992) provide a better insight into the nature of evaluation, particularly in relation to review and forward planning in education. They argue that *evaluation* is the activity which examines whether *past targets* have been met, whereas the role of *forward planning* is to *agree a strategy* for a point in the future. Consequently the education managers see a review as a link between the evaluation of the present situation with the strategic planning for the next phase. Evidence corroborating this view is provided by Baker (1991). He suggests that while assessment tends to focus on making predictions about the future and generates questions such as "where are we now and where shall we go?"; evaluation looks back at what has been accomplished and accordingly asks "Where have we been and what do we know?".

A completely separate view of evaluation has been provided by O'Keefe and O'Leary (1993), which relates to meeting the requirements of a system. They claim that an incorrect system can make costly errors or may not perform up to expectations (while are discussing expert systems their ideas could equally apply to other systems). Consequently, ensuring the quality of a system is fundamentally important. They proposed a hierarchy which showed the relationship between five aspects of the quality of a system, where each level depends upon the level below it.

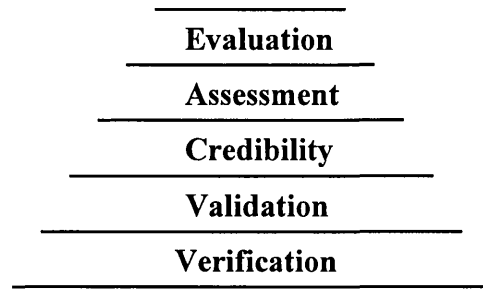


Figure 5.1: Hierarchy denoting the quality of an expert system
(from O'Keefe and O'Leary, 1993).

Their hierarchical view of system quality consisted of :-

Verification - is the demonstration of the consistency, completeness and correctness of the software (i.e. building the *system right*).

Validation - is the determination of the correctness of the final program with respect to the user needs and requirements (i.e. building the *right system*).

Credibility - is the extent to which a system is believable or how much credence can be placed in the system by users.

Assessment - which concerns other issues relating to the *fit* between the systems and the user and which are beyond the correctness of the decisions that the system makes (including the nature of the discourse between the system and the user, the adequacy and efficiency of the hardware, the quality of the implementation and the security of the system).

Evaluation - is more wide ranging issue and addresses the benefit of the system to the users, project sponsors and ultimately the organisation (i.e. is generally associated with measures of worth and value for money).

O'Keefe et al. (1993) make two important points. Firstly, traditionally different communities have concentrated on different parts of the hierarchy so that while research scientists have concentrated on the verification and validation (the so-called *V & V*), project sponsors, managers and end-users are often only concerned with assessment and evaluation of the end-product (incidentally, it is worth noting that this dipole mirrors the engineering (low-level) and managerial (high-level) views of evaluation). The authors ask us to bear in mind that any part of this hierarchy alone does not guarantee a useable system (let alone useful one). So that a poorly validated system rapidly giving wrong advice through a beautifully designed interface may be dangerous, i.e. the superficial quality of the system may induce unwarranted credibility. Secondly, that evaluations of operational systems have been based upon benefit to the organisation purely on a financial basis, such as return on investment. Since the value to an organisation is often difficult to quantify, they suffer from being accepted for development and only subsequently evaluated in an ad-hoc manner.

5.2 Why Evaluate at all?

In computer science evaluation typically takes place when new information systems are tested. This is usually to ensure the system as a whole will work and is related to two stages of the development life cycle. The first is in the system design, where each procedure is checked against its specification. When this is completed, a module (or collection of procedures) is tested. Then, a larger component of the system is tested until finally the system has been tested as a single entity. The second takes place during the implementation stage and involves operational and acceptance testing. The purpose of this is to ensure the system is able to deal with unforeseen problems arising in operation and to assure the user that the system is running satisfactorily. Kumar (1990) suggests that information systems (IS) evaluation serve to: (1) *verify* that the system meets requirements; (2) provide *feedback* to development personnel; (3) *justify* the adoption, continuation or termination of a project; (4) clarify and *sets priorities* for needed modifications; and (5) *transfer responsibilities* from developers to users.

However, in the educational field evaluation has been divided between the assessment of students, teaching methods and the selection of educational technology. Ever since the Victorian times when blackboards, lantern slides and stereoscopic viewers were used as teaching aids, teachers have had to make decisions as to what are the appropriate instructional aids for their own use. The sheer variety of educational technology has changed so much so that for example instructional technology such as computers have all but replaced radios and tapes. The task, however, remains essentially the same, that is selection of educational technology within the defining educational aims and objectives. Evaluation today is more important because of the need for accountability. As Knapper (1980) explains it "relates to the vastly increased expenditure on education in most Western nations during the late 1960s, coupled with a growing public disillusionment about the benefits of education". Hence, the current emphasis is on educational standards in all aspects. We shall return to this important change in objectives that took place in the 1960s.

Winne (1993) cites a seminal paper, by Ralph Tyler, which sketches six main purposes of educational evaluation studies:

- "● to assess the effectiveness of a program periodically so that needs for improvement could be identified,
- to verify whether hypotheses upon which a program was based had proved tenable,
- to generate information for better guiding people in a program,
- to assure those involved with a program that it was worthwhile,
- to enhance public relations, and
- describe more clearly the directions in which a program was moving."

Is there a need to evaluate educational systems? Well, as we have already seen researchers such as Nwana (1990) have pointed out there are numerous examples of intelligent tutoring systems (ITSs), (in various subject domains), being produced in academic environments. In spite of this, however, detailed research by Bento (1990) also points out that there are very few ITSs in operational use, either in the educational or training environment.

Therefore of paramount importance to evaluate, since it would be beneficial for:

1. Educational technology *designers* (and others in the development team) who may want to *improve* upon their products.
2. Other *researchers* who may wish to *learn* about the particular educational technology.
3. *Educators* and *trainers* who are likely to use it to *select* suitable educational technology for their purpose.
4. Educational technology *end-users* who would like an *input* into the design process.

5.3 Formative versus Summative Evaluation

The most pervading debate within the area of evaluation studies is the one between *formative* and *summative* evaluation. Romiszowski (1981) relates that it has become fashionable to speak of two different types of evaluation. The term formative literally means "establishing the form of a system's components" whereas summative denotes "summing up an existing system". Bento (1990) describes the former as collecting preliminary data through informal interviews, surveys, and protocol analysis. Hence being ideal for assessing at the design stage. While the latter is the assessment of the pedagogical performance of the whole system or simply a component in the target environment.

Knapper (1980) has pointed out that these two aims of evaluation need not be incompatible, but often in practice they are. This is because, most educationalists use summative assessment to provide information about an *end product*; whereas formative assessment focuses on the *process of teaching and learning* as it develops over time and requires more frequent intervention. Formative assessment would normally expected to take place in parallel with the learning. For instance Knapper describes that "In the heyday of programmed textbooks (the early 1960s) it was supposedly prescribed practice for the

fly-leaf of the book to contain details about how many times the material had indeed been tested in this way and subsequently modified." The reason behind this was that once a record is kept of every student's response to each part of the programme, this record of responses could be analyzed to determine which steps or components produce errors and misunderstandings. This could then be used to modify the programme.

As we can see now that O'Keefe and O'Leary (1993) hierarchical view of a system suggests that formative evaluations are usually the prerogative of research scientists, while end-users are only concerned with summative evaluations. Obviously, on some occasions it may be difficult to distinguish between the two types evaluation, especially if the evaluation studies are ongoing. So Hopkins (1989) made a useful distinction between the two types of evaluation in terms of the users. He states "I have noted that when the cook tastes the soup it is formative evaluation and when the guest tastes the soup it is summative."

5.4 Models of Evaluation

The field of evaluation has developed rapidly since the 1960s, giving rise to a number of different models of the evaluation process and the evaluator's role within it. However, educational evaluation has a longer history and any debate on curriculum development or evaluation must start with Ralph Tyler in 1949. Hopkins (1989) relates that Tyler became dismayed by the sloppy approach to curriculum development that he witnessed in the United States in the 1940s. Consequently "Tyler proposed as an antidote a systematic and beguilingly simple approach to curriculum planning based around four questions:

1. What educational purposes should the school seek to attain?
2. What educational experiences can be provided that are likely to attain these purposes?
3. How can these educational experiences be effectively organized?
4. How can we determine whether these purposes are being attained?"

His approach can also be expressed in an even more simplified form:



Figure 5.2: Tyler's curriculum model (from Hopkins, 1989).

Morris (1990) explained that "Under Tyler's direction, the evaluation staff...developed a theory of evaluation which was simple, powerful and so thoroughly integrated with the process of curriculum construction that it blurred the boundaries between development and evaluation. For Tyler education was essentially about changing the behaviour patterns of pupils; thus curriculum could be constructed through the specification of desirable behaviours which were stated as objectives to be achieved." Clearly Tyler's work was extremely simple and consequently suffered from a number of limitations, for example totally ignoring individual abilities and qualities.

However Tyler's approach has been enormously influential, so much so that many claim that it is the only way to develop curricula. The *objectives model* also had a profound impact on curriculum evaluation, leading to a very instrumental and outcome-oriented form of evaluation. Hopkins (1989) even suggests that the Tyler tradition influenced the first British publication in the field of curriculum evaluation (for a more detailed account of Tyler's legacy refer to Morris, 1990).

Subsequently a large number of evaluation designs came into use in education. These *models* reflected a particular evaluation method or an approach to a specific problem. Hopkins (1989) refers to Lawton's six *overlapping* curriculum evaluation models as:

- (1) The classical (or agricultural-botany) research model.
- (2) The research and development (R and D) (or industrial/factory) model.

- (3) The illuminative (or anthropological) model.
- (4) The briefing decision-makers (or political) model.
- (5) The teacher as researcher (or professional) model.
- (6) The case study (or portrayal) model.

The *agricultural-botany* method involves the assessment of different forms of instruction applied to supposedly similar classes of students. The basis of this model is the comparison of groups of students who, rather like groups of plants, are tested before the experiment (just as plant seedlings would be weighed or measured) and then submitted to different instructional methods (the agricultural equivalent of different fertilizers). At the end of the experiment, they are tested again to see which group has shown the greater amount of learning (or growth in the case of the seedlings). The numerical results that can be obtained from this type of study are apparently objective and capable of generalization to other situations. Therein lies the danger, since the method relies on the assumption that the groups of students being compared are exactly the same in every way, except for the different instructional methods to which they are being subjected. This may be true for plant seedlings, but it is obviously not true for people. No two people react in exactly the same way to any given situation, as they will have a unique combination of previous knowledge, cultural and social background, personality, etc.

The *research and development* model is essentially based on Tyler's objectives method and the approach poses the following question "Is the instruction method doing what is demanded of it?". In order to determine this, an evaluator would translate agreed aims into specific, measurable behavioural objectives. Then set up a way to assess performances after following appropriate experiences. The advantage of this systematic approach is that it enables improvements to be made after judging performance in terms of the levels of achievement set. The problem here lies in completely specifying the objectives of a programme. People are not predictable units and consequently no simple process of defining objectives and testing to see if these have been achieved will suffice.

A quite different approach is *illuminative* evaluation, which has been advocated by Parlett and Hamilton (1972). They proposed a framework which relates to research in social anthropology, psychiatry, and participant observation research in sociology. The stimulus for this alternative approach also determines its purpose, namely the need to broaden the base of traditional evaluation and relate it more realistically to the actual context of the teaching and the learning process, with all its complexity and subtlety. They describe it as a way of taking account of the wider contexts in which educational programmes function. Its primary concern is with description and interpretation rather than measurement or prediction. Consequently, the idea is to give evaluation a humanistic rather than a scientific base. The result is not a closely defined sequence of operations to be followed, but a general strategy with each investigation treated as a separate case with techniques adapted to suit its particular requirements.

MacDonald's democratic evaluation is an approach which exemplifies the *decision-makers* model, because it pursues a political analysis of evaluation (highlighting the relationships between interests, paradigms and methodology). As Morris (1990) reports that "MacDonald has provided us with a political classification of evaluation in terms of the selection of roles, goals, audiences, issues and techniques. He describes three types of evaluation - bureaucratic, autocratic and democratic - each of which have a distinctive political orientation. The tripod of bureaucratic, autocratic and democratic evaluation is set out as ideal types against which we can compare intended or actual evaluative studies....The typology attempts to provide a political conception of evaluation and outline the principles for its moral justification. It is thus a system of classification with a moral imperative but it also provides a political language with which to describe evaluation." Therefore this approach argues that though the evaluator should not make decisions, they have a political role by providing the information upon which decisions are made. The difficulty with this model is that evaluation is taken away from the origins of the work and instead is given over to *stakeholders*, with their subsidiary interests.

The *teacher as researcher* model is associated with the work of Stenhouse (1975), who wrote a critique of Tyler's objectives model. Stenhouse argued that the objectives model led to a mistaken view of education and knowledge. Morris (1990) indicates that

"Stenhouse argued that the objectives model attempted to improve educational practice by increasing clarity about ends but, 'even if this was logically justifiable' in terms of knowledge - and it is not - there is a good case for claiming that it is not the way to improve practice." The argument was there could be no curriculum development without teacher development and that the most effective way of improving practice was through teachers criticising practice. Or as Morris says "What he offered was a strategy based on teacher as researcher, rather than external evaluation as an instrument of curriculum renewal and quality control." This approach is more often commonly used as *action research*, a term used for an investigation which aims to address specific issues of classroom practice on a small, local scale. Obviously, since the work is prompted by resolving a local problem in a specific setting, it cannot said to be an evaluation carried out in any systematic and principled way. Moreover, this approach relies solely on the teacher as an expert, not only to locate the problem, but also to judge how it is to be solved.

The *case study* model is one which has been used widely in the social sciences, from history to anthropology. Bawden (1990) refers to "the value of case studies for looking at 'exemplars' - 'information rich' cases, from which a lot can be learnt. For example, in seeking to improve a social programme, it is best to look at clear examples of success, failures and indeed drop-outs. Good case studies can be far better than statistical generalizations, in giving greater personal understanding of a situation. It is, of course, vital to avoid inappropriate generalizations from case studies, and a careful 'extrapolation' is what is required." Critics of this approach argue that since a specific area for study has to be selected and it is difficult to cross-check this information, there is always the danger of distortion. Clearly we could go on to describe other models however, suffice it to say that evaluation is an activity which carries no agreed-upon definition, however, for further information the philosophical and ideological differences of this area, it is worth referring to Worthen et al. (1987). They also provide a good understanding of why a definition of evaluation is so important in the decision-making process during educational evaluation.

5.5 A Brief History of Evaluation

Some practitioners, in the general field of evaluation, have said that their profession dates as far back as the ancient world of Greece and China. As we have already indicated only in recent years, namely the 1960s, did evaluation research see a dramatic increase in the use of its methods. Bawden (1990) sees the reason for this increase "particularly marked in the USA, as a response to, and a component of, the massive social programmes, with a high degree of innovation, launched by Presidents Kennedy and Johnson (the 'Great Society'). Commentators described evaluation research at that time as 'a growth industry' and 'the liveliest frontier of American social science'". These large social programmes fuelled the need for information to guide policy decision-making ensuring accountability. One legacy of those times is that evaluation research is still dominated by the work done in the United States.

Herman et al. (1987) relates how a variety of models were advanced during the late 1960s and early 1970s in response to this huge growth in social programmes. They observe that there were two sorts of models that appeared at that time. The *first models* emphasized experimental methods, standardized data collection, large samples, and the provision of scientific, technical data. The authors reason that these models also reflected general optimism that systematic, scientific measurement procurement procedures would, as they put it "deliver unequivocal evidence of program success or failure". Furthermore this stress on *hard data* (i.e. empirically based data) would result in a rational and sound basis for planning and managing all these programmes. Indeed, so great was their confidence in their thinking at that time, that "Clear cause-effect relationships, it was assumed, could be established between programs and their outcomes, and program variables could be manipulated to reach desired effects. This first wave of models - based initially on goals and then later on decisions, issues, or problems - provided decision makers and policymakers with aggregated, standardized, and technically rigorous information about prespecified program processes and outcomes." Clearly these models were unbalanced and later they became criticised for being superficial and lacking sensitivity to individual programme variations.

As a result the *second wave* of models appeared; these were typically sensitive to unique aspects of the programmes and to the people involved. In fact Herman et al. (1987) states that "Rather than assuming that social programs were discrete and easily prespecified in terms of process and outcomes, these models acknowledged that social and other programs often are complex, amorphous mobilizations of human activities and resources that vary significantly from one locale to another, embedded in and influenced by complex political and social networks. Rare is the program, according to these model builders, which exists in hermetically sealed isolation, perfectly appropriate for scientific measurement and duplication." Although the authors acknowledged this second wave of models emphasised the qualitative and naturalistic methods and were able to accommodate the weaknesses of previous models. It also was true that the lack of generalisability to other such findings left a distinct disadvantage for high-level decision-making and accountability. More recently there has been a feeling of pessimism, querying whether these evaluations had any impact at all. Driven on by a concern to improve the chances of useful findings emerging from evaluation, a new generation of model builders modified their approach to include "socio-political and other factors beyond technical quality which were critical to an effective (i.e., useful) evaluation process." A brief description of some of these models have been given by Herman et al. (1987):

Model	Emphasis
Goal-oriented evaluation	Evaluation should assess student progress and the effectiveness of educational innovations.
Decision-oriented evaluation	Evaluation should facilitate intelligent judgements by decision makers.
Responsive evaluation	Evaluation should depict programme processes and the value perspectives of key people.
Evaluation research	Evaluation should focus on explaining effects, research identifying causes of effects, and generating generalizations about programme effectiveness.
Goal-free evaluation	Evaluation should assess programme effects based on criteria apart from the programme's own conceptual framework, especially on the extent to which real client needs are met.
Advocacy-adversary evaluation	Evaluation should derive from the argumentation of contrasting points of view.
Utilization-oriented evaluation	Evaluation should be structured to maximize the utilization of its findings by specific stakeholders and users.

Table 5.1: Some models of programme evaluation (adapted from Herman et al. 1987).

Herman et al. (1987) add that each model comes from a set of interests and circumstances and therefore has its own strengths and weaknesses. Moreover although each has its own emphasis, they are not mutually exclusive, and we would add that each is distinguished more from what it doesn't stand for than what it advocates. Some of these models are discussed in more detail in Worthen et al. (1987). They examine how each approach has been used and the strengths and limitations of each one, from both a theoretical and practical perspective.

This fleeting glimpse should at least have provided the impression that even though the field of evaluation has a relatively short history, its deficiencies cannot be the lack of approaches. In fact taking the education domain Morris (1990) has indicated little had changed from the use of such approaches, "It could be argued that evaluation in both Britain and the USA has had little direct impact on education policy, and that where it has had impact it may well have served to complicate rather than rationalize the policy-making

process. The large-scale educational evaluations of the 1960s and 70s were generally unsuccessful at providing unequivocal evidence about programme effectiveness." His reasons for this disillusionment rest on the fact that ultimately the evaluation had to be justified in terms of providing the type of information needed by policy makers to produce change. Since this is political in nature, the success of the evaluation has to be judged in terms of whether it effects policy outcomes. In that context, it competes against a variety of methodologies and political interests, clearly it has not come off best.

An alternative explanation is offered by Worthen et al. (1987), who cites the lack of maturity as a cause, "In a young field there is inevitably a good bit of conceptual floundering as notions are developed, circulated, tried out, refined, and challenged by new alternatives. Until a solid knowledge base begins to guide practice, any new field is likely to be guided by the positions, preferences, and polemics of its leaders. Some may argue that this is inappropriate. But the point is that it is also inevitable, for no new field or discipline is born full grown." Certainly this does explain the large repertoire of approaches to evaluation. A number of writers including Morris (1990) have claimed that since the 1960s the field of evaluation has settled down a bit and tried to concentrate on finding the most appropriate methodology, rather than devising a new one or arguing between different ones. This pragmatic view, sometimes known as responsive evaluation, takes its direction from the matching of evaluation methods to specific evaluation situations and questions.

5.6 The Style of Evaluation Chosen for this Study

From this discussion of evaluation it is clear there are several principal approaches to evaluation, depending on the questions being sought, the ideological stance which is taken and the circumstances of the evaluation. Knapper (1980) suggests that evaluation studies in instructional technology can be broadly classified into four categories: *simple descriptive studies*, *comparative studies*, investigations of *different modes of implementation* of an instructional system, and *overall programme assessments*. We shall describe each of these in turn, however it is worth noting that some types of research

approach are more common than others. For instance, the most popular is the comparative study where a particular instructional technology is compared with conventional instruction and the least favoured approach involves a general assessment of entire instructional programme. In addition, Knapper reports that not all types of instructional technology have been studied to the same extent, "Instructional television has, for instance, received a vast amount of attention in the research literature whereas there is much less interest in, say, teaching via radio".

The *simple descriptive* studies constitute a straightforward account of a particular instructional method and as Knapper notes it is "often written by someone who has been closely involved with the innovation". Information which is generally reported is about such important factual questions as what the technological system comprised of, how it was organised and implemented, how many students took part, what cost were involved. The information may be scientifically *objective* in terms of the questions and the evaluator may include subjective judgements about the value of the exercise.

Comparative studies are the most common type of investigation and they compare the effects of two or more different methods of instruction. This is either carried out by the manipulation of variables (i.e. experimental designs) or simple measurement (i.e. correlational designs) or even a combination of both. Knapper relates that a popular approach "is to use the medium of instruction as the independent variable, with student attitudes, learning gain, time taken for learning, instructional costs and so on as dependant variables. Rather few of these studies have attempted to compare different types of instructional technology, the great majority concentrating instead on comparisons between a particular innovative approach to teaching and so-called 'conventional' or 'traditional' methods - meaning instruction that is teacher-dominated and largely oral".

Research on *different methods of implementation* is about how the instructional technology may be applied most efficiently. Although it focuses upon a single instructional innovation, it does go beyond the simple description of its organisation and effects. An example of this style is given by Knapper involves whether "programmed instruction more effective for students working as individuals or in pairs? If in pairs, then how should the

students be matched: on the basis of similar abilities, similar personalities or some other criterion? How is student performance with a particular instructional medium affected by the attitudes of the teacher in charge of the course (positive or negative)? What sorts of delivery format promote most interest: for example in the case of radio instruction, how do straight talks compare with dramatisations?". This type of investigation can be carried out either by fairly orthodox experimental designs or alternatively by correlational designs.

A more uncommon approach is to study the complete performance of educational innovations over time and in a variety of situations and is comes under the heading of *general programme assessments*. This type of investigation uses a variety of techniques including simple description, correlational studies, case studies and experiments. The advantage of this approach according to Knapper "is their very comprehensiveness, which permits the researcher to 'stand back' and get a fuller picture of the strengths and weaknesses of the instructional approach concerned". It is also clear why this way of evaluating is so rare - the costs both in terms of time, effort and money are likely to prove prohibitive.

We decided to conduct the present study in the comparative style of evaluation for a number of reasons. Our overall goal is to encourage the development of a number of measures which assess the value and effectiveness of operational teaching systems. This means finding out whether a system can teach students *effectively*, what *makes* it a good system and whether this could be *improved*. Examining the different styles of evaluation that are available for instructional technology, we can see the simple descriptive study will only provide illustrative information regarding a particular system. As for different modes of implementation, this is not useful since it is solely about customising a system for use. The overall programme assessment seemed to be a robust approach and one which could yield useful results about the strengths and weaknesses of the system. However, it is outside the resources and time limit allowed for carrying out a study under the conditions of a PhD thesis. Consequently, comparative research appeared to be ideally suited to the conditions and requirements for our evaluative study. This approach would enable the effectiveness of a system to be measured in relation to other forms of teaching (thereby establishing a benchmark). Measures of learning could be taken before and after both

methods of instruction. In addition a number of other dependent variables could be obtained in order to make sure that if any other variable were to change in the study (apart from the intended method of instruction) then this would be accounted for. Another advantage of this approach is that we can be fairly certain that the effects of long-term exposure to the system (e.g. boredom, prior usage and maturation) are limited.

5.7 Key Points

- Evaluation is the activity which examines whether past targets have been met (Rogers et al., 1992) and as a consequence it is perceived as a link between the present situation and the next phase. Most educationalists refer to two types of evaluation: (i) formative and (ii) summative. The former takes place early in development and the latter provides information about the end product.
- The field of evaluation dates back to the 1960s, when the increase in social programmes required these programmes to be reviewed, accountable and ultimately controlled. This also facilitated the interest in development of models for evaluation. The first wave of models was based on “hard data”, these were gradually replaced by the second wave which were based on qualitative and naturalistic data collection. Since then models have matched the evaluation method to the specific type of investigation and situation.
- Knapper (1980) suggested that evaluation studies in instructional technology can be broadly classified into four categories: (i) simple descriptive studies, (ii) comparative studies, (iii) different modes of implementation and (iv) overall programme assessments. The comparative style of investigation was chosen for this research for the reasons provided.

CHAPTER SIX

QUALITY SYSTEMS

"Quality....you know what it is, yet you don't know what it is. But that's self-contradictory. But some things *are* better than others, that is, they have more quality. But when you try to say what the quality is, apart from the things that have it, it all goes *poof*! There's nothing to talk about. But if you can't say what Quality is, how do you know what it is, or how do you know that it even exists? If no one knows what it is, then for all practical purposes it doesn't exist at all. But for all practical purposes it really *does* exist. What else are the grades based on? Why else would people pay fortunes for some things and throw others in the trash pile? Obviously some things are better than others....but what's the 'betterness'?...." [Pirsig, 1974, p.187]

6.1 Introduction

Over the last three decades we have experienced a phenomenal growth in the use of information technology. This rapid rate of technological development could not have taken place without the advent of microprocessors. While it is impossible to predict exactly what the future rate of expansion will be, it is clear from the preceding decade that a number of trends are likely to influence further utilisation of this technology: processor speeds, storage capacities, integration, complexity and quality of software. In terms of computer hardware great improvements have been made, so much so that, for instance Norris et al. (1993) reports that a personal computer today is more powerful than an average mainframe a decade ago and data storage costs have fallen by several orders of magnitude over the last ten years. They foresee that this sort of progress could well be repeated over the next decade. At the same time, we have seen the costs in software development rise dramatically. To give some idea of this increase Sommerville (1989) indicated that costs are growing at a rate of 12% per year. Due to this, and other reasons such as its irreplaceability, computer software is coming to be seen as the critical resource in the information age.

It was recognised in the late 1960s that a large number of major projects were late, cost much more than originally predicted, were unreliable, difficult to maintain and performed poorly. Software development was in a crisis situation. The causes were thought to be related to building large and complex systems. Cox (1995) relates how a new expression came into use during this period, "the NATO Software Engineering Conference of 1968 coined the term *software crisis* to indicate that software was already scarce, expensive, of insufficient quality, hard to schedule, and nearly impossible to manage." Even though a number of improvements, such as education of development staff, increased use of tools, and changes in approaches, have taken place the software crisis is still with us.

Yeates (1991) pointed out that quality in software development is as important today as it was in the past. "In 1988 in the UK, Price Waterhouse announced the results of a study that revealed that £500m a year is wasted by users and suppliers of computer-based systems as a result of quality defects in software. A US Government Accounting Office report shows a similar position in the US. This study reported on a number of federal software projects costing a total of \$6.3m. Half of these were delivered but never used, a quarter were paid for but not delivered and only \$300,000 worth was actually used as delivered and even then abandoned or substantially reworked. It has been estimated that two thirds of all software development errors result from inadequate work at the requirements analysis and design stage, and the other third during the coding stage. Also, for every hour spent correcting a problem at the requirements stage, it takes 100 hours to find it at the testing stage." (1;14;12;9;5;14;1;18) In the light of this why has quality, in particular, taken on such a significance? To answer this question we need to know more about quality. Therefore in this chapter we discuss what is meant by quality, how to ensure that a quality system is produced and especially how to measure for quality in a system.

6.2 What is Quality?

The first pertinent question to ask is, what is quality? Many people misunderstand this word, therefore it is worth clearing this up. The common definition given by The Oxford Paperback Dictionary is:

"quality.....a degree or level of excellence,.....a characteristic, something that is special in a person or thing,"

However this is *not* what managers mean by quality. Instead a better definition is the one used by ICL, quality is "Fitness for purpose". Joseph Juran, a quality management guru, is credited with coining this phrase. Another definition of quality, which was provided by Dr Edward Deming (generally acknowledged as introducing *quality management*) explains its wider implications

"A predictable degree of uniformity and dependability at low cost and suited to the market."

This has come to mean that quality is giving complete satisfaction - providing customers with exactly *what* they want, *when* they want it and at a *price* they can afford. This can be illustrated with reference to the manufacturing industry, where the ideas of quality originated. To exceed a requirement may appear to increase quality to the user. But it means that the producer has spent more time and effort than required. For example, should a manufacturing company produce screws for a car plant then its quality is maintained, providing the supplier keeps the screws to the size and shape required by the customer. However, if the machine makes too many screws which exceed the size tolerance levels, i.e. too big or too small then they are rejected as waste. They are poor quality because they are, just as any consumer rights group would describe, "not fit for the purpose described". One should now be aware that quality is not an ephemeral goal but is

intimately related to customer requirements. Stebbing (1989) stated that "Quality, like beauty, is in the eye of the beholder. What is considered by one person to be of good quality, could be considered by another to be of poor quality, and vice versa." One might say, this is fine for manufacturing but, what lessons does this hold for software development?

In software engineering the cyclical, waterfall and 'V' life cycles are well known. In this context the objective of quality is to "get it right first time". In other words, to make sure the end product meets all the original requirements and/or specification, therefore avoiding time-consuming rework and reducing costs. This means quality is applicable to every phase of software development (especially in drawing up a detailed specification). In implementation, for example, the code should be *readable*, *maintainable* and *correct* thus meeting design requirements and passing tests. The significant point here is that quality for any company means assessing its customer's needs and expectations and making sure they can fully satisfy them. There are three quality functions which follow from this statement:

- Quality Management (QM) - the setting of general quality policy and standards. A concept which is related to Total Quality Management (TQM), that is ensuring the company recognises its customer needs and expectations and also satisfies them at the first attempt. It must be an approach carried out throughout all of its department and activities.
- Quality Assurance (QA) - retrospectively checking the quality of a product, usually after it has been built (however, quality planning begins at the earliest stage in the process).
- Quality Control (QC) - the application during a project of a set of procedures and techniques, whose purpose is to ensure a particular quality of the product and satisfies a given goal or objective of the user (this is one aspect of quality assurance).

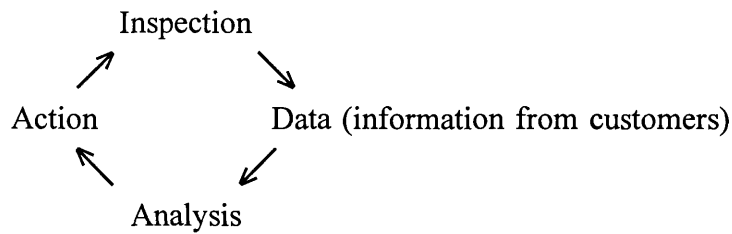


Figure 6.1: Quality Control.

Yeates (1991) cited Philip Crosby who "has promoted the idea of a quality culture and framework. He established the 'four absolutes of Quality Management.' These are

- Quality is defined as conformance to the customer's requirements and not as 'goodness' or 'elegance'.
- Quality is implemented through the prevention of defects not through post-manufacturing inspection.
- The performance standard must be 'zero defects' and not 'that's near enough'.
- The measurement of Quality is the price of non-conformance."

Here are a few examples of Quality Assurance across a software development life cycle:

Phase	QA procedure
Requirements	Prototyping
Design	Inspection
Code	Walkthroughs
Test	Review/Checklist

Table 6.1: Quality assurance procedures.

Quality assurance is not just about dealing with software faults, it is also concerned with reliability, as well as other attributes such as maintainability, portability, readability, etc. Two things should be now quite apparent, firstly, quality is not just an inspection activity at the end but an integral part of the product and therefore considered vital. But, it also must be plain that quality is a sea change in management, or as Stebbing (1989) affirms, "quality assurance is a philosophy of total integration of the business to achieve the required result."

6.3 Quality versus Productivity

Furthermore, it should now be understandable why an emphasis on quality is superior to an emphasis on productivity in the project setting. It is easy to see that there is huge gulf between the two concepts. Fundamentally this follows from their definitions. Whereas productivity simply relies on rolling out more products in a shorter time, quality stresses the importance of meeting customer requirements, which means information plays a key role. In the past the emphasis has been on productivity rather than quality. It can still be seen during system development where something like 80% of the errors are made in the requirement phase. It is further disheartening since approximately 70% of these errors are solved in maintenance. This is a huge burden on the project in terms of time and cost that cannot be overemphasized.

An example of the benefit of quality over productivity was given in the October '91 issue of American's business magazine, Fortune. In a survey of the 100 fastest growing companies, Michael Dell (founder of the Dell Computer Corporation) was asked why Dell was rated so highly. He replied that it was largely due to two managerial practices, often touted but seldom practised:

"No.1: putting the customer first - listening, understanding, serving....It responds to customer wants and needs rather than letting managers arrogantly insist they instinctively know what's right."

and

"No.2: moving with relentless speed and adaptability. Unencumbered by commitments to the old ways....newer companies have the edge here."

Given this, why have project managers ignored quality and insisted on productivity? Lock (1988) provided a common reason "When a project is in danger of running late, there is always a danger that corners will be cut at the expense of accuracy or quality. A high incidence of modifications arising from design errors could be the direct result of undue haste or crash action. In normal circumstances it would be usual for all drawings to be checked, and for design calculations to be verified (sometimes by means of a prototype...)." The scope of this discussion restricts further details. However a good collection of findings for runaway projects is given by Mousinho (1990) and, of course, Brooks (1975).

The second point to be made is that, traditionally productivity improvements have been sought by manipulating three main areas: the employee, the task or the environment. As a consequence the product loses its priority. In contrast, quality turns this notion on its head and focuses exclusively on the product (we shall see in the next section this has a number of beneficial side-effects). Even now, when quality is taught in project management it is largely confined to a small portion of testing (see for example, Bentley 1982). This criticism is not restricted to the Computer industry as Farr (1984) observed "Why is it that Japanese products, which were so recently regarded as 'cheap oriental trash', are now regarded as world beaters in both quality and reliability?....In many ways the disappointing performance of British industry is largely due to our attempts to adopt the American system of scientific management, often referred to as the Taylor System. This System relies on the creation of a management elite in which instructions are passed to a supposedly subservient labour force. Under such regimes almost all responsibility is removed from the foreman and passed to specialists such as Quality Engineers, Work Study Engineers, Production Engineers and Inspectors." (4;1;18;5;20;8;15;19)

Quality is not a small function of the total operation. Quality in project management is a *key consideration* and should be regarded throughout the entire project life cycle, impacting on the determination of requirements, design right through to maintenance and ongoing support. Indeed Stebbing (1989) dispelled the myth that quality has only a specific role and asserted "Although a quality assurance programme will include quality

control and inspection, both these activities form only a part of a company's total commitment to quality."

Thirdly, the implications of implementing quality are in complete contrast to productivity. Quality is a long term aim which entails more initial costs, more commitment and more changes. Whereas productivity is a short term solution, which has minimal costs, and inherently manages with few real changes. Robson (1982) further highlighted this difference, "The Taylor system was often instrumental in achieving substantial increases in productivity, in the short-term at least, and spread like wildfire, becoming the prevailing method of organising work throughout the developed world....Over the subsequent....years the Human Relations movement, with names like Mayo and Maslow at the forefront, ranged itself against Taylor's Scientific Management, and a great debate started as to which way was most likely to produce positive results....In Japan this message was avidly received and developed into a Japanese version of what is normally known as Total Quality Control. In Japan TQC meant every employee in the organisation having a responsibility for quality, and this, of course, meant that a massive education job needed to be done. Remarkably, one of the ways this was tackled was by a series of radio programmes, aimed mainly at first line supervision, which provided a frame work of quality control knowledge." (5;23;8;15;14;5;22;5) There are many consequences resulting from this course of action. For instance, two of these are retraining (to provide the appropriate skills) and participative style of work (encouraging responsibility for the quality of the product).

Finally, returning to the software crisis we can clearly agree with Cleland (1990) when he asserted that "A project management philosophy that focuses on quality throughout the management of a project is one of the best ways that the project manager can meet the challenge of bringing the project in on time, on budget, and within technical performance objectives."

6.4 Total Quality Management

Japan stands out as a country that has successfully taken the quality philosophy to its heart. Are there any reasons for their success and can we learn them? Hutchins (1984) stated "Contrary to popular opinion, the success of the Japanese has little to do with their being workaholics, as they have been recently described, or with their mysterious oriental culture. Prior to 1949 the Japanese were operating none of the systems which have subsequently proved to be so effective." Ironically as Yeates (1991) pointed out that quality in Japan originated from an American expert, W. Edward Deming. "He spent much of his time in the immediate post 1939-45 war years in Japan where he found that managers were receptive to his ideas about quality. This was in contrast to western managers at that time whose perception was that quality increased the cost of a product and adversely affected productivity. Following Deming's advice, Japanese companies took a number of important actions. Firstly, they invited customers into their factories and worked with them to improve quality. Secondly, the responsibility for quality was given to all line managers who worked to their own clear quality objectives. Finally, responsibility for quality was delegated to all levels of staff. The emphasis was clear. Effort should go into controlling the production process rather than merely inspecting the end product." Possibly the most important lesson Hutchins has for us is that Britain's forgotten asset is the co-operation between manager and employee, shown most clearly in the employee's lack of authority or training and therefore a tendency to be ill-equipped to handle the problems which arise.

Robson (1982), on other hand, affirmed that participation and ability are emphasized in quality cultures, most notably expressed through Quality Circles. Quality circles are a way of getting a team to contribute ideas to solve quality problems. The circle should decide for itself which problem to consider, although management usually retains the right of a polite veto. They usually made up of between three and fifteen people. As Gillies (1992) suggested "A quality circle is a group of workers who are asked, not told, to join. They should be an overall supervisor to coordinate the whole quality circle programme throughout an organisation. Finally, management must be committed to the programme.

Whilst they retain the right and obligation to manage, they must not reject recommendations without good reason or they will strangle the idea at birth." A good account of the experience of Quality circles in Japan and America is given by Shea (1991).

Quality then is a goal worth aiming for. However, the introduction of quality can present certain difficulties especially when carried out properly. For instance it should never under any circumstances be seen as the latest management gimmick. It is also worth bearing this in mind when considering what Stebbing (1989) pointed out about quality, "it is not a panacea for all ills'. Quality assurance will not cure everything but it will go a long way towards getting things right first time, every time." This is amply illustrated with reference to the growing number of companies, some computer-related, which are undergoing certification and registration under BS 5750. This is a third party check by the BSI (British Standards Institute) that a supplier is demonstrating adherence to certain quality standards. Following successful assessment the organisation is included in a register of approved companies, which is open to public examination. Therefore, it has two main purposes. Firstly, externally, it demonstrates to the customer the company's commitment to quality. Secondly, within the organisation, it aims to review the services available and consider which are best suited to its requirements and marketing needs.

BS 5750 has five guidelines, but essentially it is only divided into three distinct parts (from which you can choose which best suits your company):

Part 1 : Specification for design, production, installation and servicing.

Part 2 : Specification for production and installation.

Part 3 : Specification for final inspection and test.

The International Standards Organization (ISO) have also developed a series of quality regulations, which correspond to above part numbers ISO 9001, 9002 and 9003

respectively. While the content of ISO and BS are identical, slightly different titles are used.

The pros and cons of BS 5750 come under two headings, namely the introduction of BS 5750 and the end result of certification and registration. Taking the former first, there can be a large number of difficulties arising from implementing this process. Firstly, the company must be willing to get the commitment of its managers and understanding from its staff. Then there is the problem of a lack of experience of quality systems in some companies. Added to this is the obvious extra burden of BSI, thoroughly checking every procedure. Underlying this is the time and cost needed for the implementation.

So what are short-term benefits? Well, there will be an immediate identification of strengths and weaknesses of the company. Communication will be easier partly due to the identification of customers and suppliers (not just outside the company but also within). Following from this, there will be an increased control and visibility from improved knowledge of responsibility and outcome of work. Simple things emerge, such as training will be easier since everything is written down.

Overall, there are many long-term advantages and disadvantages of quality systems. Briefly, BS 5750 only demonstrates that quality will be fully considered by a supplier. It does not, however, guarantee that problems will not occur or that a product will be totally successful. It only increases the likelihood of success of the product. There is little evidence that product price is reduced by a quality system. On the other hand, the advantages of a quality system does mean bad practices are identified and rectified. It reduces the time and cost of rework, correction and product support. Purchasers also favour suppliers registered with BS 5750. This is probably because the end-product is easily a saleable one, since the requirements have been automatically met. Furthermore, if any additions are required, it can be easily supplied since the product is very maintainable.

An example comes from a February '92 issue of Computing about OKI, a Japanese company operating in Scotland. They put so much stress on producing reliable products,

delivery on time and focus on the customer that "finished goods also rigorously checked for defects but OKI's definition of a defect is far-reaching. If someone forgets to sign a check sheet, for example, that is regarded as a defect. This sort of attention to detail is intended to remind people about the quality process." Clearly this level of scrutiny is rare, but not unknown in a company with a quality culture and illustrates the pride taken with the product.

6.5 Measuring for Quality

In this research, evaluation of developed systems means that quality is likely to play an important part. Although the subjects of evaluation and quality assurance are related there does not seem to be many references to such an overlap in the literature. The main focus for the following discussion is how we can measure if a quality system has been produced. Measurement of quality is seen as an extension of quality control since, as Moller and Paulish (1993) report, "Metrication provides visibility on the entire software development process, and it thus reduces its apparent complexity. This takes away some of the 'mystery' and 'wizardry' of software development, and makes it more understandable through its similarity to other industrial development processes." Although, as we shall see it is not that simple to obtain such measurements.

6.5.1 Standards and Benchmarks

An area worth examining, although briefly, is software standards. A standard may be commonly described as a prevailing example by which the product may be tested or measured. Standards are important in all sorts of products since they provide directions for improvement. Indeed Sommerville (1989) argued that one of the most important roles of a quality assurance team in a large software project is the development of appropriate software standards. He went on to divide software standards into two types, process and product standards. Process standards refer to how the software process should be conducted, for example setting out how design reviews should be conducted and

documented. Whereas product standards, which dominate this set of discussions, set out attributes which all product components should exhibit. An example of a product standard is a programming language standard which sets out how a language should be used. Therefore, we can see that general quality standards such as BS 5750 and ISO 9000 certify the process of producing software and not the products which result. An example of a product standard is the BSI certification mark, known as the Kitemark. A product with this mark confirms that it has been manufactured to a published specification and undergoes repeated independent testing, surveillance and inspection; in this case for safety, quality and reliability.

Traditionally technology standards serve two purposes according to Evans et al. (1993), on the one hand they provide a consistent and conventional environment i.e. *sameness for its own sake* (examples include arrangement of instruments in an aircraft, frequency of the electrical supply and indicating the side of the road which we drive), while on the other hand they guarantee presence of a certain property (examples include electrical or mechanical safety), this is usually an issue of fit. Moreover, in the latter the objective is to ensure uniformity within limits or at least a minimum level. The authors put forward the following motivations for standards: fit (to make communication possible), safety (to reduce the risk of injury to human beings), protection (from poor quality), cost savings and reduction of inconvenience. They add that most of the benefits of programming language standardisation falls into the last category.

Out of all the information technology areas, Evans et al. (1993) proposed that software standardization is perhaps the most complex. They explain that it faces particular problems because of two main reasons, firstly since it is soft and not hard and secondly because software possess a flexible nature, "Software standards have to deal with abstract rather than concrete entities (though some make matters worse for themselves by muddling up abstractions and realizations on actual hardware). Possibilities are infinitely variable, limited only by the range of inventiveness of those concerned....The flexibility of software encourages customized solutions and attitudes opposed to the idea of standardization. The apparent freedom from limitations makes it harder for people to accept that some self-discipline is desirable. Arguments against standardization are advanced such as that

standards impose constraints when there is no need, that standards inhibit progress, that standards are extremely expensive to develop and implement - all based on the idea that this limits the benefits of adaptability without compensating benefits in return."

A standard, which is often mentioned in computer science, is the *benchmark*. It is a simplified comparison often used as a way of indicating how quickly computers perform. French (1984) defined benchmark as "a standard program, or assembly of programs used to evaluate hardware or software. For example, a benchmark might be used to:

- a. compare the performance of two or more different computers on identical tasks,
- b. ascertain if the configuration provided by a manufacturer performs according to his claims,
- c. assess the comparative performance of two Operating Systems." The reason for referring to this in passing is that even though it is a so called *hard* standard, different benchmarks are needed for different situations.

6.5.2 Quality Metrics

Software metrics is the area to do with measurement and prediction of the quality and complexity of software products. The need for software measurement is evident today as it was three decades ago. In 1992 an article, in Computer Weekly by Stalley, drew attention to the need for software metrics, "General managers want to run the ruler over their IT operations, and in response, systems managers are being tempted to turn to some form of software metrics in order to demonstrate effectiveness....It will be able to discover and demonstrate the areas of strength and weakness, and have in place a system of metrics which can be used to drive a process of continuous improvement."

Although Fenton (1991) pointed out that software measures have had a long history, until recently this area of software engineering has been largely neglected. The reasons are

twofold, firstly what little measurement is performed "is done infrequently, inconsistently, and incompletely". Secondly, as with standards, such measurements are not easy to define. Norris et al. (1993) reiterated that, on the face of it, software engineering is merely a new activity to which traditional skills and standards need to be applied. To some extent this is true, but they argued that components of software operate with less visibility and are less tangible than other engineering objects, consequently it is difficult to create reproducible measures or standards for. Another difference is that, apart from words and pictures, there is no tangible end product. They end by saying "Given that software engineering is only just over 30 years old, it is hardly surprising that a rich and mature literature, from which engineers can learn, is yet not in place. The move from art to science may have started for software but it still has some way to go." Alternatively there are others, such as Yeates (1991) who argue that since quality is itself subjective it cannot be measured at all. Although even he admits adopting such an argument means that quality is never really seriously tackled and avoids the issue. So attempts must be made to define the quality of software.

Norris et al. (1993) made two key points about software engineering which we take up further:

- "● The detail of software matters. One mistake can make the difference between success and failure and it is difficult to state what is required with both clarity and precision.
- The intangible nature of software make it *inherently* error prone. The customers rarely get exactly what they think they have asked for.

Put these two factors together with the sheer complexity of modern systems and it is no great surprise that traditional engineering project management does not transfer altogether successfully." The intention of this work is to develop system metrics which manage to incorporate both the actual detail and the user's point of view, so that it should point to how software should be developed, that is this approach should enable designers to produce the desired result. From what we have learnt, an important component in developing a quality system is meeting the customer requirements. Cooper and

Kleinschmidt (1991) cited a study which supports this assertion, the study sought to identify those characteristics that were common to new product success. This extensive study of 567 successful innovations by Myers and Marquis found most successful products were market-derived (market pull) ventures and that only 21% were technology push. Consequently, the way to tackle a major improvement in computer-assisted instruction is to discover measures which increase the success of the system.

Fenton (1991) proposed that in any software measurement activity there are three classes of entities whose attributes may be measured. Like Sommerville (1989), Fenton also talked about processes and products but he added another entity, that of resources which he described as "inputs to processes" for example people, software tools and hardware. Furthermore he stated that each of these three classes may be subdivided further into internal and external, where:

"● *Internal attributes* of a product, process, or resource are those which can be measured purely in terms of the product, process, or resource itself.

● *External attributes* of a product, process, or resource are those which can only be measured with respect to how the product, process, or resource relates to its environment."

There seems to be a close analogy of these two types of attributes with formative and summative evaluation. Measurement of internal attributes usually takes place early on and with the designer of the software, whilst the measurement of external attributes often takes place at the end, typically by the software manager or user. Fenton suggests that while internal attributes can be measured directly and are relatively easy to measure, external attributes are indirect and most difficult to measure. Moreover there does seem to be intuitive connection between the internal structure of software products and its external product attributes. Indeed, built into most software quality models, there is an assumption that *good* internal structure leads to *good* external quality. However there have been very few scientific attempts to establish specific relationships between the two.

Since our objective is to derive quality metrics for computer-assisted instruction and the ISO formal definition of quality is: "The totality of features and characteristics of a product or service that bear on its ability to satisfy specified or implied needs." [ISO, 1986], i.e. in this case, how well the system meets the user requirements. Consequently, our intended measures must fall into the category of external product attribute, since it is a type of measure dependent on entities additional to the product in question. Some of the common external attributes, which users would like to know, are identified by Fenton: reliability, maintainability, usability and portability, of these software reliability seems to be especially important.

Furthermore, Fenton (1991) described how in most quality models these attributes exist at a high level hierarchy and must be decomposed into lower-level attributes, normally called quality criteria, to make them meaningful. He added that "In many cases the quality criteria are internal attributes, such as structuredness and modularity, which developers believe should impact on external quality attributes. One further level of decomposition is required in which the quality criteria are 'associated' with a set of low-level, directly measurable attributes (both product and process) which are normally called '*quality metrics*'." Further information on hierarchical models of quality and the interrelationship between quality criteria can be found in Gillies (1992). Lastly, it is worth noting that since our discussion has focused on software considerations, we have not mentioned quality in education terms. This does not mean that aspect is irrelevant, but has had a limited influence on the current work.

We intend to derive a set of metrics for our own work, adopting the Software Quality Metrics (SQM) model as described by Basili (1995), "The SQM approach was developed to allow the customer to assess the product being developed by a contractor. Thus, the object of the SQM evaluation is the final product for the purpose of assessment from the point of view of the customer. In this case a set of factors is defined on the final product. These factors are refined into a set of criteria, which are further refined into a set of metrics. The models are [pre] defined and the user selects the particular set of factors and criteria of interest." Essentially this is a model for defining external product quality attributes. We are convinced that quality metrics is a useful approach for complex and

intelligent systems, indeed a recent group of researchers, Juristo, Mate and Tovar (1997) are building a measurement environment for knowledge-based systems (KBSs). They have based their work on the Goal Question Metrics (GQM) model of quality. We are in agreement with Adelman (1992) when he cites Grady and Caswell, “A metrics program is NOT a software *project* - it is part of the *process of management* of every project....[it] is the starting point for instituting....quality improvement.”, certainly metrics have a role to play in computer assisted instruction, what that role is - that’s too early to tell.

6.6 Key Points

- Poor quality systems are caused by the failure to deliver software “fit for the purpose” and this wastage often means re-work and unsatisfied customers. As a consequence, the goal of quality is to “get it right first time”, in other words to make sure the end product meets all the original requirements and/or specification, therefore avoiding time-consuming rework and reducing costs.
- Quality achieves this objective by making customer requirements explicit and by keeping control of this measurement. However, quality assurance is not just about dealing with software faults, it is also concerned with reliability, as well as other attributes such as maintainability, portability, readability, etc. Indeed, quality is not just an inspection activity at the end but an integral part of the product and even acts on the development process.
- One important way of making the requirements explicit is to convert them into “quality metrics” (thereby making them visible, tangible and measurable). Fenton (1991) describes two basic types of measurements that can be made: (i) internal measures, based upon the “white box” view of the software representation and (ii) external measures, based on the “black box” view of the software. The former is a vantage point looking out from inside the system whereas the latter is looking in from the outside. We have adopted the Software Quality Metrics (SQM) model for our study because it is an approach used on final products to derive external product quality attributes i.e. suitable for operational systems.

CHAPTER SEVEN

RESEARCH DESIGN

7.1 Introduction

This chapter covers the design of the research study, that is the selection of evaluation technique, the arrangement of students and the practical implementation. This chapter also provides an overview of some of choices available and presents *how* the research design was chosen and on *what basis* the design choices were made. It was deemed necessary to elaborate on the variety of research designs available because these are rarely, if ever, discussed in a computer science discipline. Moreover since this work involved evaluation, it is worth noting that scientific research and evaluation are not necessarily compatible (for further information refer to Kidder, 1981 who provides a good contrast and comparison between these two approaches).

7.2 Determining the evaluation

The first step in the evaluation process was to decide what type of evaluation should we be aiming for. This involved determining what was expected and what was the purpose of the evaluation? In addition we needed to know what information is expected from the study. Our overall goal was to examine a fully functioning computer-assisted instruction (CAI) system within an operational environment. The purpose of this research was to develop a way of assessing the effectiveness of CAI systems. Two aims came out of this goal. The first aim was to find out whether the system was at least as effective as conventional teaching. The second aim was to determine which key features or characteristics (i.e. criteria) of the system are significant in assessing the quality of a teaching system (where quality is defined as *fitness for purpose*).

7.3 Approaches to Scientific Research

There are two basic approaches to scientific research: *ex post facto* (also known as correlation research in the social sciences) and experimental. *Ex post facto* literally means "from what is done afterward" and as Kerlinger (1973) pointed out it refers to something done after an event with a retroactive effect on the event. While a *true* experiment is an investigation where the investigator manipulates and controls one or more independent variables and observes the dependent variable(s) for any accompanying effect. In addition, *true* experiments are marked by the ability of the investigator to freely assign experimental groups and experimental treatments. Since the independent variable(s) have already occurred in the *ex post facto* research, the investigator cannot manipulate or assign experimental groups or treatments. In practice, this distinction between a *true* experiment and *ex post facto* design breaks down.

7.3.1 Experiments

Experiments are tests designed to arrive at a causal explanation. The most typical and certainly simplest experimental design consists of two groups, where individuals are randomly assigned to either an experimental (i.e. treatment) group or a control group. Both groups are then exposed to *exactly* the same conditions, with the exception of the experimental group which receives some *special treatment*. Subsequently, the investigator measures both groups to assess whether the groups differ. Since both groups have been treated equally, we can be quite sure that any difference(s) between groups are solely due to the *special treatment* implemented on the experimental group.

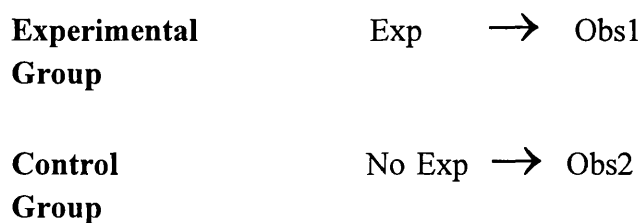


Figure 7.1: The Simple Experimental Design.

where *Obs1* and *Obs2* are measurements taken after each condition, *Exp* is exposure to treatment and *No Exp* is no exposure to treatment.

As we indicated earlier, the most important advantage of the experimental method is that it allows for direct cause-effect conclusions from the data. If there is proper control in a particular experiment, that is variables other than the independent variable are kept constant and it is observed that the dependent variable changes as the independent variable changes, therefore the investigator can conclude that the relationship between the independent and dependent variables is causal. In non-experimental research methods, causal relationships cannot be drawn.

Another advantage of the experimental approach is that it can be repeated, not only by the researcher who devised it, but by anyone who is able to follow the same procedure. Repetition of an experiment under controlled conditions reduces doubts if similar observations are made each time it is performed. Agreement among different observers is an important advantage.

Three important limitations exist in experimentation. The first is that experimentation cannot always be used. For example, areas such as human eugenics and selective breeding are not experimentally explored for ethical reasons. A second important limitation of the experimental method is that it is artificially arranged by the scientist. There are two possible consequences of this. One is that the investigator acts on judgement which may lead to error. Sometimes the investigator might select a variable which may be believed to be significant but in reality is not. The other result is that conclusions from an experiment may be limited to the artificial experimental situation - they may not apply to natural situations, or even to other experimental situations. The third limitation is that experiments could interfere with the very factor they are attempting to measure. In many situations, it is extremely difficult and sometimes impossible to control all irrelevant variables.

7.3.2 Basic Experimental Designs

Some of the most important decisions relate to the way in which individuals are assigned to the different conditions. The manner in which this assignment takes place results in the three basic designs available in the literature:

- (1) Independent measures designs;
- (2) Repeated measures designs; and
- (3) Matched subjects designs.

7.3.2.1 Independent Measures Design

For this design individuals are randomly allocated to one or other of the experimental conditions, with the single constraint that there will be equal numbers in the two groups. This usually done by using random number tables.

The allocation might be as follows:

A	B
S1	S3
S2	S6
S4	S7
S5	S8
S9	S10
S12	S11
S13	S15
S14	S16

Table 7.1: Example of subjects in an independent measures design.

Here *S1* stands for individual one, *S2* for individual two, etc.

In this experiment each individual will provide us with a single score for purposes of analysis, i.e. the total number of individuals will be the same as the total number of scores. This kind of experimental design is also known as *between-subjects* or *unrelated* design, because the comparison is *between* two groups of subjects whose scores are *unrelated*.

But there is one crucial disadvantage about using different individuals to do different experimental conditions. There may be all sorts of individual differences in the way different individuals tackle the experimental task. As a result, all these individual differences in performance might be affected by irrelevant variables which have nothing to do with the independent variable, which the investigator is intending to manipulate.

For this reason it is usual to allocate different people at random to experimental conditions. The reasoning is that, if individuals are randomly allocated to experimental conditions on a chance basis then people of different ages or abilities are just as likely to be found in all the experimental groups.

This design would be used, for example in the following circumstances:

- (a) Where there is an adequate supply of subjects for the experiment to allow the randomization procedure to be effective (say 20 to 30 in all).
- (b) When the stimulus material used in the experiment has to be the same for each level of the independent variable but cannot be exposed twice to the same subject.
- (c) When the experimenter involves only one subject.

7.3.2.2 Repeated Measures Design

In this design, a single individual appears under both of the experimental conditions. Thus for the same number of scores as the other two designs we need half the number of individuals.

The allocation is as follows:

A	B
S1	S1
S2	S2
S3	S3
S4	S4
S5	S5
S6	S6
S7	S7
S8	S8

Table 7.2: Example of subjects in a repeated measure design.

Here *S1* stands for the same individual under both experimental conditions, *S2* stands for a second individual, who also appears under both conditions, and so on.

This design is also known as a *within-subjects* or *related* design, because the comparison is *within* the same group of individuals, the scores from each individual being *related*.

From the point of view of eliminating individual differences this is clearly a better experimental design than random allocation of individuals, since it uses the same individuals for all experimental conditions. But, there are special problems relating to the order in which each individual performs under the two experimental conditions.

So that while this design ensures individual peculiarities get equalised out over all conditions. There are special problems relating to the order in which each individuals performs under the two experimental conditions. These difficulties arise because other factors, such as the type of and the ordering of treatments and so on, may play a part in

affecting the dependent variable. These so-called *carry-over effects* (from one treatment to the next) need to be taken into account when designing such an experiment. For instance any order effects can be counterbalanced, half of the individuals are given treatments A followed by B, whereas the other half are given the reverse order i.e. B followed by A. The point is to try and eliminate all possible irrelevant variables which might be affecting the individual's performance.

This design is useful in cases where order effects are known to be symmetrical and therefore they are amenable to the counterbalancing technique.

7.3.2.3 Matched Subjects Design

In this design, individuals are matched in pairs and the members of the pairs are allocated randomly, one to each of the experimental conditions. The matching can be performed in terms of a third variable which the investigator believes is likely to affect scores on the dependent variable. Hence it is also known as matched-pairs or related design.

The allocation might be as follows:

A	B
S12	S11
S21	S22
S31	S32
S42	S41
S51	S52
S62	S61
S71	S72
S82	S81

Table 7.3: Example of subjects in a matched subjects design.

Here *S11* stands for individual 1 in the first matched pair and *S12* stands for individual 2 in the first matched pair. *S21* stands for individual 1 in the second matched pair, and

so on. The decision about whether 1 or 2 within each pair is allocated to condition A or B is on a random basis.

As with the independent-subjects design, each individual will provide us with a single score for analysis, the total number of individuals again being the same as the total number of scores.

It may seem that matched-subjects designs do indeed combine all the virtues of using same subjects and different subjects. Because the groups consist of different people it is possible to use the same materials and there is no need to worry about order effects. Yet, because sets of matched individuals are allocated to experimental conditions variability due to the characteristics of different individuals should be much lessened.

Unfortunately, matching subjects has problems of its own. Often it is just not possible to find subjects who are matched on all relevant characteristics. It would be rare indeed for an experimenter to know beforehand all the individual variables which are likely to have an effect on people's performance. Moreover, in most kinds of experiments are likely to have small numbers of people and therefore, with the exception of large-scale research, it is not possible to select groups of subjects to be matched on important variables.

In order to match subjects it is necessary to equalize all possible relevant variables, so that matched subjects can be treated for the purposes of the experiment as if they were indeed the *same* person.

In practice, these designs are rare - they are mostly used in studies of twins - even in these studies the method is open to the sort of criticisms we have made.

7.3.3 Other Experimental Designs

Up to now we have been talking about situations in which there is only one experimental group and one control group from whom the independent variable is withheld. This is the

simple or classical experimental design, more commonly known as either the *control-group* design or the *randomized two-group* design. The point about this control-group design is that it should result in both groups being equivalent in every way except for the independent variable, since individuals are randomly assigned to the groups. It has already been emphasised that experiments are designed to arrive at a causal explanation. The objective of an *ideal* experiment is to tightly control the study so that the effects on the dependent variable can be assessed and the range of possible explanations as to the cause of any observed effects can be reduced to a minimum. However, there are alternative designs which attempt to improve upon this control. We shall look at a few of these designs and comment about their advantages and shortcomings.

One way of achieving better control is a slight variation on the repeated measures design. Here, individuals are measured on the dependent variable before the treatment and then again following the treatment. Consequently the effect of the independent variable on the dependent variable can be compared with the individuals own baseline. This *before-and-after design* gives a good control over individual differences, because such individual differences are present both before and after the independent variable is introduced. This design is also known as the *A-B-A within-subjects experimental design* because the first A is the baseline condition without the independent variable, B is the condition with the independent variable and the last A refers to a final measure following the independent variable.

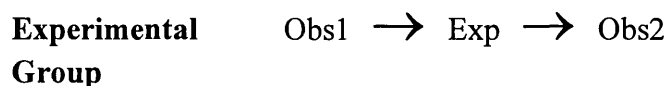


Figure 7.2: The A-B-A Within-Subjects Design.

where *Obs1* is measurement taken before treatment, *Exp* is exposure to treatment and *Obs2* is measurement following treatment.

At face value, this design would seem ideal for controlling extraneous variables since each individual acts as their own control. However, Kerlinger (1973) argued this design does

have some disadvantages. Firstly, it possible that the pre-test could sensitise the individuals and therefore the post-test measures could be affected purely by this reaction. Secondly, another set of extraneous variation is history and maturation. Clearly depending on the period of time many things can occur between pre-test and post-test, other than the effect of the independent variable. Furthermore, other more general events which come under the category of maturation can occur, the ones connected with change or growth in the individual. Another less well-known drawback is the phenomenon of the regression effect. Scores of tests change, it is recognised that on retest the average score *regresses* toward the mean.

A different variation on the previous design is to randomly assign individuals to two groups (experimental and control groups) which undergo pre-tests and post-tests. Unsurprisingly this design is known as *before-after two group design*.

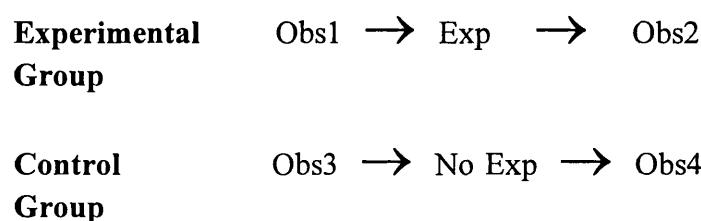


Figure 7.3: The Before-after Two-group Design.

where *Obs1* and *Obs3* are measurements taken before each condition, *Obs2* and *Obs4* are measurements taken after each condition, *Exp* is exposure to treatment and *No Exp* is no exposure to treatment.

As we have seen previously, pre-tests can provide a more sensitive test of the treatment by letting each individual serve as their own comparison. Pre-tests can also provide other advantages, firstly they can provide a check on the randomisation and let the experimenter see whether the groups were equivalent before treatment. Moreover, if the groups are not equivalent on the pre-tests, the investigator can adjust the post-test measures to provide a fairer test of the treatment.

As before, pre-tests are subject to the same criticisms. Thus the pre-test may sensitize individuals and bias their post-tests. Should this arise in this design then it should happen for the experimental and control groups alike, their post-test scores should be equally elevated or depressed and pre-testing alone would not be a rival explanation for a difference between Obs2 and Obs4. If however, the pre-test affects the treatment group differently from the no-treatment group, this would appear as a difference on the post-test scores and would be indistinguishable from a difference produced by the treatment alone.

It should now be clear that an experiment is no better than its controls. Investigators must be very careful to control their experiments adequately. Since there are so many extraneous factors that can influence the dependent variable, it is important to control for as many of these as possible so that when interpreting experiments rival explanations are ruled out. Kidder (1981) pointed out that ruling out rival explanations depends on three things: "(1) the design of the study, (2) the experimenter's knowledge of what different events (such as intrasession history) the subjects may have experienced, and (3) the pattern of results and whether that particular pattern could plausibly be explained by some events other than the experimental treatments".

One experimental design, which has been proposed as strong and aesthetically satisfying in terms of controls, is the *four-group, before-after design* (also known as the *Solomon design*). This design combines features of the classical experimental design and the before-after two-group design. Four groups are needed because in addition to the treatment and no-treatment groups, there are pre-tested and nonpre-tested groups. The main advantage of this design is that the investigator can test decisively whether the post-test differences were caused by the treatment, the pre-test, or the combination of treatment-plus-pre-test.

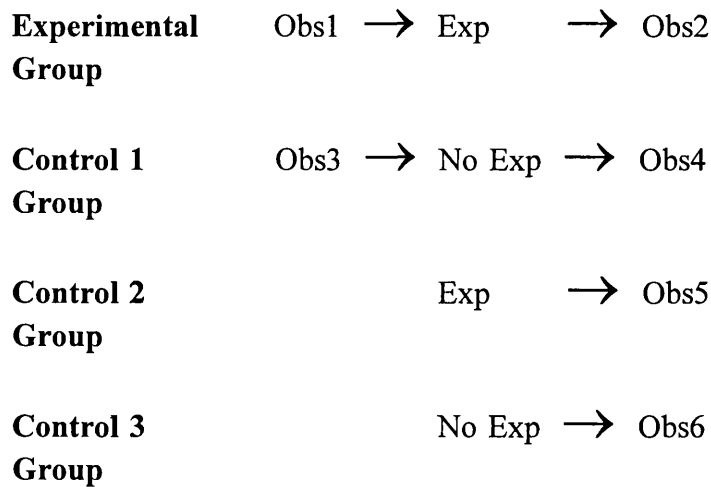


Figure 7.4: The Four-group Before-after Design.

where *Obs1* and *Obs3* are measurements taken before each condition, *Obs2*, *Obs4*, *Obs5* and *Obs6* are measurements taken after each condition, *Exp* is exposure to treatment and *No Exp* is no exposure to treatment.

The value of this design is derived from the combination of the two designs (mentioned previously) into one strong design. Its virtue derives from several reasons: the comparison between experimental group and control 1 group is well satisfied, as well as the comparison between control 2 group and control 3 group, the randomization of individuals assures statistical equivalence of the groups and history and maturation effects are controlled by the experimental and control 1 groups. Pre-test sensitisation and its interaction effect is controlled by the first three conditions. The fourth condition enables possible temporary effects between pre-test and post-test to be controlled for.

Though this design is certainly a strong one, and is held in high esteem within the research community, it does suffer from two sources of weakness. Firstly, this design produces practical difficulties which centre around running two simultaneous experiments and the problem of obtaining a large group of individuals of the same kind for all conditions. The second obstacle is statistical and arises because, although there are four groups, there aren't four complete sets of measures. This means that it is difficult to

obtain one overall statistical approach. Kerlinger (1973) argued that while this design offers a number of advantages, on the other hand the sheer complexity of the design means it is doubtful whether it can be used routinely and consequently should be saved for occasions when already tested hypotheses can be tested again with rigour and control.

Until now we have examined designs where there is one experimental group, with one (or more) control group from whom the independent variable is withheld. However, there are occasions when several experimental conditions or treatments are required - for example when different levels of computer-assisted instruction are provided, i.e. minimum, medium, and maximum. Designs with two or more independent variables are called factorial designs. The benefit of these designs is that two or more independent variables can be manipulated at the same time. Therefore it becomes possible to see whether the combination of two independent variables is different from what we would expect if we simply added the separate effects of each variable. Such combinations are known as *interaction effects* and are amenable to examination in the Solomon design.

A group of designs known as *time designs* are commonly used in the educational field to study development and growth. These are longitudinal studies of individuals over different points in time. One such design might be:

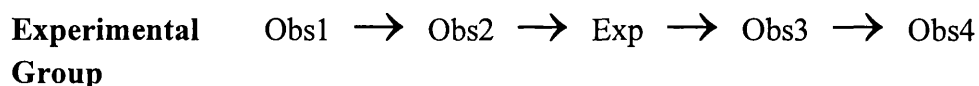


Figure 7.5: Example of a Time Design.

Note the similarity to the before-and-after design where a group is compared to itself. However, though reactive measurement effects can be accounted for, longitudinal or time studies do suffer from one difficulty - they cannot control for the natural effects of time, i.e. individuals do not stop growing and learning for the convenience of research. Since the influence of time has a confounding effect and can prove especially troublesome it can be difficult to avoid spurious results. A number of common tactics to reduce this problem are suggested by Kerlinger (1973). Firstly to add more control groups, another is to add more time observations and a third is to add more experimental conditions.

7.3.4 Correlational Studies

Correlational studies are concerned with identifying, or measuring the strength of, association or correlation between variables. Correlational studies cover a number of different forms of research, including natural observation, survey and case study (all of which is discussed later on). Common to all of them is that the investigator is simply examining what is happening or has happened *without* deliberately producing changes. Therefore unlike experiments, the investigator cannot manipulate independent variable(s) and observe the resulting effect on dependent variable(s). In correlation studies the investigator simply passively observes the relationship between a pair, or pairs, of variables. As a consequence inferences about the causal relationship between the variables cannot be made and all one can say is that there *is a* relationship between the variables. Indeed, even dependent and independent variables are not usually distinguished prior to the study. As Robson (1983) pointed out:

"In natural observation, all variables are free to vary. Because of this, inferences which we make from natural observations and other non-experimental methods suffer from a serious defect. We can never be sure that changes in one particular variable occur as a result of changes in a second particular variable. In other words, from natural observation we can make no definite statements about causal relationships between variables. It is always possible that a third variable is causally related to both of the first two and has produced the relationship observed between them."

Occasions when correlational studies are appropriate, are either when nature has already implemented the independent variable (i.e. *ex post facto* research) or when, although it might be possible to manipulate a variable, it is not ethical to do so. In other words, experimentation can be seen as a special form of observation, i.e. controlled observation. A basic difference between experiments and other forms is that the experimental design advocates deliberately manipulation of one variable and at the same time controls all other variables, so that they do not affect the outcome.

As we have seen this poses a heavy limitation on the experiment, since it distorts the investigation situation into an artificial one. Hence it has been argued on a number of occasions (for example, Elmes et al. 1985 and Kerlinger 1973) that the true experiment circumscribes the generality of the experimental findings. Another argument against experiments is that either extrapolation or interpolation from the results is restricted by the homogeneity of the individuals (as we have seen they are required to be that way because of the need to control for individual differences). This issue is commonly referred to as *validity* (more specifically ecological validity). Since the design of correlational studies is to capitalise on variations that occur naturally, it is not surprising that correlational findings are more open to generalisation to other situations. A related benefit is that the resultant model is more likely to reflect the *real world*. Robson (1983) contended that this disadvantage of lack of control is more than enough compensated by the considerable benefits of the naturalness of the setting, "a warning against over-optimism and against the rejection of non-experimental wisdom. There was a wave of enthusiasm for experimentation in education in the 1920s. This was followed by a wave of pessimism and disillusionment. The advocates of experimentation assumed that progress in teaching methods had been slow just because scientific method had not been applied. When the experiments proved to be tedious, equivocal, difficult to replicate and to accord with common sense, then disillusionment and rejection of experimentation took place. The justification for the experimental method does not lie in the fact that it is a panacea, but that it is the best method we know for getting at cause-and-effect relationships." Robson goes on to say that there are many situations where direct experimentation are not possible. Indeed perfectly respectable sciences such as astronomy and geology have to rely almost exclusively on natural observation.

7.3.4.1 Naturalistic Observation

Naturalistic observation refers to the systematic observation of a certain event or phenomenon in the environment as it occurs naturally, without any intervention on the part of the investigator. The researcher avoids manipulation if they believe that it would adversely affect the basic characteristics of the particular phenomenon.

Observation, or naturalistic observation, has always been a tool of science. In fact, the first requirement of a good scientist is that they must be a good observer. Observation is a well respected and scientific method for accumulating information. The procedures of observation are simple but they must be followed rigorously. The observer simply observes the people and environment they are interested in and records each and everything with absolute precision. However, observation must be directed. The observer must have a question or set of questions in mind before they undertake observation. There are millions of observations that one can make and therefore it is important that the observer focus their attention on a particular aspect of the situation. Typically, the observation is directed only by the events required to prove or disprove hypothesis that is to be tested. However, it is worth noting that the kinds of questions which are asked will constrain to some extent what kinds of answers will be produced. In fact, rather than answering any questions off one's own bat, careful observation should enable us to pick out regularities and get some idea what the appropriate variables might be.

Sperling (1982) gives an example of one type of observation in a different discipline which illustrates the method's benefits. "Perhaps the simplest observational technique of all is the *field study method*. This consists simply of making observations with no attempt to control conditions or to get the co-operation of your subjects. In fact, field studies are usually most successful when the subjects are unaware of being observed. Great use of this method was made in England during the Second World War. To determine how well morale was holding up, trained observers successfully 'spied' on the people. The field-study method differs from personal experience in two ways. Actual counts are made, and more people are encountered than most individuals can meet."

Observation however, is not only carried out by social scientists and ethologists, but it is also a method commonly employed by systems analysts during the fact finding part of their systems investigation. Its drawbacks include an expensive dedication in time and effort, problems in remaining unobtrusive, huge amount of information filtering and the danger of only providing a snapshot of a changing scene. However, as Clifton (1986) remarked it can prove to be invaluable method for systems analysts, "When done discreetly, this method yields information unobtainable through other methods".

Furthermore, this approach can be useful in cases where the investigator is either unfamiliar with the situation or has an idealistic impression of what is actually taking place.

7.3.4.2 Case Study

The Case study involves extremely detailed observations and reflections on individual cases. As Bell (1987) has indicated "The great strength of the case-study method is that it allows the researcher to concentrate on a specific instance or situation and to identify, or attempt to identify, the various interactive processes at work. These processes may remain hidden in a large-scale survey but may be crucial to the success or failure of systems or organizations."

The researcher's task is to identify common and unique features and to show how they affect the implementation of systems. The researcher's methods involve close observation and questions, usually in the form of an interview. Interviews may be structured or unstructured, depending on how much is known about the phenomenon. Structured interviews involve responding to a number of pre-formulated questions. An unstructured interview is one in which the interview is guided by the responses and the researcher's hypotheses concerning the underlying rationale of these responses. Case studies can also be accompanied by compiling archival documentation relevant to a particular problem or issue.

Critics of the case-study approach draw attention to the fact that it is not usually possible to generalise the results and question the value of single events. While supporters of the case study, such as Bell (1987), assert that if the research is carried out systematically and critically it can provide "a three-dimensional picture and will illustrate relationships, micropolitical issues and patterns of influences in a particular context."

7.3.4.3 Survey research

Another important method of research is the survey, since it enables the investigator to gather information on a large sample. As Bell (1987) indicated "The aim of a survey is to obtain information which can be analysed and patterns extracted and comparisons made." The census is the most common example of a survey. In surveys, individuals are asked to reply to a series of questions or items. The investigator wishes to learn an opinion, attitude, or belief on a particular issue. However unlike an experiment, the investigator records information without actually doing anything, in fact the objective is to try as much as possible not to exert any influence. Indeed, Bell (1987) asserted that "Information can be gathered by means of self-completion questionnaires (as in the case of the census) or by means of questionnaires, schedules or checklists administered by an interviewer. Whichever method of information gathering is selected, the aim is to obtain answers to the same questions from a large number of individuals to enable the researcher not only to describe but also to compare, to relate one characteristic to another and to demonstrate that certain features exist in certain categories. Surveys can provide answers to the questions What? Where? When? and How?, but it is not so easy to find out Why? Casual relationships can rarely if ever be proved by survey method. The main emphasis is on fact-finding, and if a survey is well structured and piloted, it can be a relatively cheap and quick way of obtaining information."

Inevitably, there are problems in the survey method. Great care has to be taken to ensure that the sample population is truly representative. If a decision is made to carry out a survey then there needs to be consideration of what characteristics of the total population need to be represented in the sample, so that it is possible to say with fair confidence that the sample is reasonably representative. In addition, when conducting a good survey leading questions must be phrased in such a way so they don't suggest *correct* answers. Moreover, all the questioning must be carried out with the same eagerness ensuring that more thorough information is not elicited from one type of person than from another. All of these problems can be sorted out by weeks of planning, reading, design and exploratory piloting.

7.3.5 Quasi-experimental designs

Quasi-experimental design is a type of research rarely considered in the scientific literature. Quasi-experiments lie somewhere between *true* experiments and correlational studies. They are useful on occasions where the independent variable is not under the investigator's control or where it cannot be directly manipulated. For instance, when the variable *occurs naturally* (e.g. tornadoes) or when it is *ethically unsound* to alter the variable (e.g. gender of the subject). As a consequence the quasi-experimental design makes comparisons between treatment and non-treatment conditions using non equivalent groups or using the same individuals prior to treatment. For instance comparing the reading ability of speed readers with average adults. There are countless numbers of quasi-experimental designs that an investigator can construct (a few examples are provided by Kidder, 1981).

One advantage of quasi-experiments is their focus on naturally occurring independent variables (which have a high degree of intrinsic interest and important practical implications). In addition, they take advantage of observational and correlational procedures and combine them with the power of experimentation. Quasi-experiments can be less intrusive than *true* experiments. However, since this research is conducted with what is given, i.e. without proper comparison groups or random assignment of treatments, it is likely to forfeit control (lack of internal validity). This non random sorting makes it difficult to disentangle the treatment effects from other rival effects, particularly selection, but on the other hand they give this design an element of external validity.

7.4 Benefits and Drawbacks of Experiments and Correlational studies

The following table illustrate the benefits and drawbacks of every approach discussed:

Research method	Advantages	Disadvantages
Experiments	Establish cause and effect relationship.	Artificial results. Ethically unfeasible for all occasions. Difficult to control all extraneous variables.
Naturalistic Observation	Naturalness of setting and results. Information not obtainable by other means.	Time consuming. Huge amount of information to record and analyse. Problems in remaining unobtrusive.
Case study	Allows understanding of individual process unlike large-scale research.	Inability to generalise results. Resource-intensive.
Survey	Relatively cheap and quick way of obtaining information about a large sample. Enables relatively small samples to be targeted so that results can be generalized to the population.	Sample needs to be representative. Questions need to be carefully developed.

Table 7.4: Comparison of experimental and correlational studies.

7.5 Choice of Research Design

This passage describes a summary of the decisions which were made in selecting the research design for this study. Primarily our chosen design had to take into account the aims of the research, which were to compare how effective a particular CAI was in

comparison to conventional teaching and to determine the features significant in assessing a teaching system. Due to a time limitation of 6 months to a year to organise the study as well as carry it out, it would have been difficult carrying out a case-study which would have meant a large amount of time conducting a lot of interviews with individuals. Furthermore we would have obtained results that would have been difficult to generalise, therefore case-studies were eliminated.

The next choice was to collect a large amount of information using an observation. Although this is good method for finding out variables in a realistic setting (especially where little is known about the particular field), it also suffers from being time-consuming. In addition, it would be difficult to establish the effectiveness of a CAI system simply by observing. We quickly realised we had to intervene in some way, in order to determine whether the CAI system introduced was better or worse than the current teaching method.

Consequently, we decided that an experimental design was the only way we could find out how effective the CAI system was. The design we started out with was a simple experimental design, with two groups - an experimental group, which would receive the CAI system and the other group (i.e. control group) who would not be exposed to it. However, this was inadequate since the control group should be the conventionally taught group, so that a comparison could be made. But how could this be achieved this in real educational institution, where experiments could not be carried out since some people would *suffer* from not receiving exactly the same teaching? Eventually, we found a way round this problem by asking the lecturers to split the group into two. So the control group would be taught conventionally while the experimental group would be taught using the CAI system. Then a test could be taken to compare the two groups. After which the experimental group could go on to be taught the course topic(s) they missed out on. This design was further modified subsequently when we decided that not only did we need a post-test but also a pre-test in order to provide an indication of improvement in learning. So the design that was eventually chosen was the Before-after Two-group design.

The final problem was how to design the study so that we were able to collect information about the student's assessments of the strengths and weaknesses of the system. While an experiment is good at determining casual relationships between variables, it does not provide a wide variety of information about *why* this is the case. Consequently we decided to use a structured questionnaire to elicit the relatively large amount of information quickly and easily. We would use the questionnaire *only* on the group which had used the CAI system. Although we would need a lot of time to develop the questionnaire, it would be relatively straightforward in collecting the results using this method.

CHAPTER EIGHT

PROCEDURE OF STUDY

8.1 Introduction

The previous chapter outlined the reasons for choosing the design for this research. It was a mixed design involving both an experimental and a survey component. This chapter gives a *personal* account (hence the use of the first person singular throughout) of how the research design was implemented in a practical study and the obstacles which ensued from this decision. The reasons for providing information about the difficulties encountered is to help other similar studies to estimate their own timetables. The purpose of the survey was to gather information from the users about the strengths and weaknesses of the system. In order for this to be done a structured questionnaire was needed to capture the multifaceted assessment from the people who had used the system. The process of creating the questionnaire is discussed later on in this chapter. The other part of the design was a Before-after Two-group experiment. This was intended to compare students who used the system with those who had been taught conventionally (i.e. experimental versus control group), but also to compare each student's performance prior to and subsequent to their differing teaching sessions. As pointed out in the previous chapter, this design proved to be difficult to achieve within an applied setting of a university. A critical problem to do with this was that although the experimental and control group would be treated differently (to be scientifically valid) however, in education this is unethical and inappropriate since one group would invariably benefit at the other's expense. In order to tackle this conflict between the objectives of theoretical research and the restrictions in an applied setting I decided to carry out the study on the experimental group before they were taught conventionally (so that their overall education would not be affected). Moreover I decided that after the complete study had taken place, I would offer *all* students a chance to use the tutoring package (thus not *depriving* the control group of any

beneficial education effects). This was also to prove problematic since I had to organise my study around the teaching schedule.

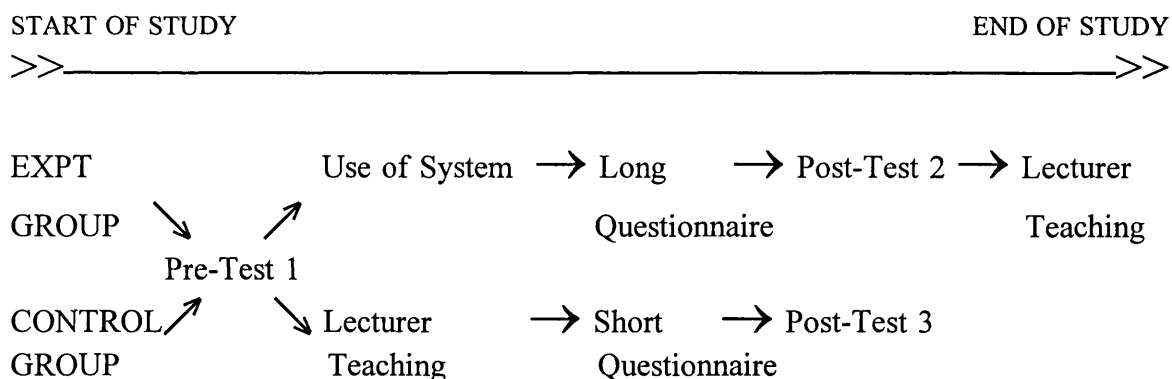


Figure 8.1: Representation of the procedure of this study.

8.2 Obtaining a fully-functional tutoring system

One of the most important but also frustrating tasks of the research was to obtain a computer-assisted instruction package which was in *operational* use within a university. The requirement for a *fully operational* system was sought in order to ensure a comprehensive evaluation and such an assessment is not possible on a pilot, a prototype or partial system. The requirement for a university system, as opposed to one in use in a school, is that the assessment by the users is likely to be somewhat easier for several reasons. One is that the questionnaire need not be so simplified (taking account of naive users). Another reason is that it is expected the participants would be more sophisticated computer users and therefore they would give a more critical response. With this in mind, I began to contact prominent researchers in the field, such as university lecturers with a CAI interest and other people who were likely to know or have access to such operational systems.

I was disappointed by the response since most did not reply, however, even the ones that did either indicated they didn't possess such a system or referred me to someone else. I

became very exasperated by the huge effort I was making trying to contact anyone with a suitable system and that I was receiving only limited help.

Finally, after a relatively long period and after making strenuous efforts to seek an *operational* intelligent tutoring system I located one at the, aptly named, Electric Brain Company (EBC) from information provided by Professor Alistair Kilgour of Heriot Watt University. The APT (or Application Program Tutor) is *the only* commercially available ITS which is in operational use in this Britain. One of the local sites, where it is being used is Thames-Valley University, Ealing. In order to obtain a copy I had to submit a proposal of my research and send assurances from my department that I was only intending to use the system for the purpose of research. These items were sent off in December '93 and in March of the following year I received the software disk and documentation.

The APT is a knowledge based tutoring system designed specifically to develop and teach courses for prospective users of computer applications. It can be described as an expert system shell which typically consists of some form of knowledge representation scheme and a ready made inference mechanism. In addition, the shell contains facilities for producing good interfaces for the application user. However there is no domain knowledge stored in this system, instead a knowledge acquisition tool helps elicit and organise the knowledge. It consists of three main system components: APTWatch, APTBuild and APTTeach. APTWatch is the knowledge acquisition part of the system in which the course designer is prompted to walk through the target application functions and a screen image is captured of the application highlighting an important aspect of the application. APTBuild will redisplay the screens of the action as they were captured, enabling the course designer to input requirements or choice of options (including background detail to support the trainee).

APTTeach uses a selection of teaching and assessment strategies for teaching procedural skills in order "guide" the trainee through each task for that course. Five different strategies are detailed in the system documentation: (i) cognitive apprenticeship - performing a demonstration of the task, then asking trainees to perform sub-part of this

task, (ii) successive refinement - makes selective passes through the course starting with the overview and moving down through the lower levels, (iii) abstraction - which enables a keyword that links across tasks to be explained fully or in summary depending if a task is completely taught, (iv) direct assessment - trainees are asked questions (including multiple choice) when the task has been taught, and (v) practice assessment - asks a trainee to perform a task which has already been taught. Trainees are required to master each task before they can proceed to the next. Courses are required to last less than 45 minutes, corresponding to between 8 and 10 sub-tasks. Typically courses can be on databases such as dBase or wordprocessing packages such as Microsoft Word. The screen dumps of the course designed using the APT can be seen in appendix U.

At around the same time, I was given an article by a colleague (Paul Wernick) about an accounting tutor. This led to a meeting with a group of people at De Montfort University about a computer-assisted instruction package they had developed to teach their accountancy students. After discussing my research aims, I was able to obtain their support and acquire a demonstration copy of their marginal (variable) costing *tutor*. The developer of the system, Ashok Patel, briefed me about the software's history (see Wilkinson-Riddle and Patel, 1993). It was part of the CTI initiative (a Government backed project, involving primarily ex-Polytechnics collaborating, to develop teaching technology). He also informed me that De Montfort University had the system on open-access to students in the college library and Middlesex University were one of a number of institutions using and testing the system on their own students. Therefore by April, I had two different educational software packages to assess. The next phase would involve searching for appropriate students to be used as subjects in the research but before this I had to prepare questionnaires and assessment tests. I choose the accountancy tutor as the starting point (since it needed the least amount of work preparing it for an evaluation).

The marginal costing system, that was used in the study, is a DOS-based CAI which uses a tutorial to introduce basic concepts of management accounting such as break-even points, unit revenue and contribution to sales ratio. Then whilst in the "interactive mode" the user is required to solve a number of randomly generated problem scenarios known as "computer generated examples" (using the newly acquired marginal costing concepts,

relevant feedback, hints and reference to the tutorial examples or their own saved examples). In addition, the user will be able to look up a graphical representation of the marginal costing data they have entered, using break-even charts, i.e. enabling them to compare the on-screen effects of alternative decisions (which is the overall objective of management accounting, namely testing what-ifs with real business data). Besides this "management tool" feature, the tutor is capable of entering a mode in which there is "automatic marking of assignments" (with no feedback given). The system was designed to be very interactive and easy to use for accountancy students in a self-paced tutorial (see appendix P for screen captured examples of this tutor).

8.3 Learning about marginal (variable) costing and testing the software

Once I had obtained the software I set about studying its functionality and getting completely acquainted with the software inside and out. At the same time since I had no prior knowledge of marginal costing, therefore made myself familiar with the concepts by reading a few books on management accounting, for example Lucey (1992) and Mott (1984). There are two major areas of accounting known as financial accounting and management accounting. The former relates to generating information for groups external to an organisation and the latter relates to producing reports internally, most notably for the managers so that they can make business decisions. Marginal costing can best be explained with reference to the two main types of costs that management have to deal with, that is fixed costs and variable costs. Fixed costs (or fixed expenses) are expenditure that remains unchanged, in total, irrespective of changes in the levels of production or sales (such as business rates, rent, and some salaries). Variable costs are those that increase as the amount of sales or volume of production increases, for example raw material expenditure or delivery charges. The key baseline for managers is determining the break-even point i.e. the level of production, sales, percentage of capacity, or sales revenue at which an organisation makes neither a profit nor a loss. In this context, marginal costing is defined as the change in total cost (i.e. fixed costs plus the variable costs) incurred as the result of the production of one additional unit of output.

One of the key concepts I learned in marginal costing was Margin of Safety. The Margin of Safety (MS) is defined as the distance between the current output and the break-even output. It can also be expressed as:-

$$\text{Margin of Safety} = \frac{\text{Profit}}{\text{C/S ratio}}$$

OR

$$\text{MS as a percentage (\%)} = \frac{\text{Profit}}{\text{C/S ratio}} \times 100$$

Figure 8.2: Equations defining margin of safety.

Where C/S is the contribution to sales ratio. If the MS is short it indicates that a large reduction in output may cause the firm to make losses. If however the MS is large, it indicates the firm can reduce output substantially without making losses.

8.3.1 Derivation of Attitude Items

I knew that a questionnaire needed to be developed that was quick and easy to fill in (since most of the target sample i.e. accountancy students, were expected to be unfamiliar with questions in this format). But also it had to gather relatively large amounts of complex data, but also be amenable to easy analysis. Once I was familiar with the relevant aspect of accountancy, I began reading the software documentation. Using the knowledge of the system, its documentation and the criteria I had been investigating, I set about generating statements for the questionnaire. As already indicated, the aim of the questionnaire was to solicit a detailed assessment of the system's strengths and weaknesses in terms of its key features. In this regard I turned to Adelman (1992) who emphasises that an effective evaluation needs to decide which questions should be answered. He says the evaluator does this by identifying *measures of effectiveness* (representing the evaluation criteria). These *measures of effectiveness* should be determined from a decision-making

perspective, i.e. what information is needed, who needs it and what stage is the system development process in, and so on. Furthermore these measures must be correlated with the overall utility of the system.

I was also influenced a lot by John Cook's (1991) MSc Thesis, especially by the way his questions were phrased. He underlined an approach which aimed to discover the users model of the system, their understanding of the available operations and their interpretation of screen information. Although, his work revolved around obtaining qualitative data using observational recordings of user and the system and opinion data (in structured and unstructured questionnaire format). He related a technique of *critical-incidents* (in which people are asked to describe the most *noteworthy* aspects of their interaction with the system) and breakdown, which are both are used in a framework of *co-operative evaluation* i.e. which considers the users as co-evaluators of the system. Accordingly, I started by examining the *specification* within the documentation, in order to determine what was claimed the system could do. Some of these design features were clear from the documentation, for instance "it assumes no prior knowledge of computing or accounting" and "it provides meaningful feedback to students". While others were less clear and had to be inferred, such as "it allows the student to solve a problem in any order" and "it is more effective learning than reading a textbook (because it is interactive)". From this careful examination of this documentation and the system I managed to accumulate 132 separate items. The next stage was to arrange these items together in related groups. It is worth noting that both Adelman (1992) and Jordan et al. (1990) represent such measures of evaluation criteria as a hierarchy, so that for example the latter argues ultimately there are three major characteristics of a successful system: functionality, ease of maintenance, and flexibility. After arranging the items together I obtained sixteen main categories (analogous to these characteristics) relating to different aspects of the system.

The next stage was to determine how to pose questions using this set of diverse features. I consulted a number of references about questionnaire design, such as Oppenheim (1986), Rust et al. (1989) and of course Bell (1987). They provided a range of alternatives from checklists to grids. However in the end, I chose to use a rating scale for several reasons,

that is, once has been explained it is quick and easy to use as well as administer, it allows the respondent to provide a more sensitive measure to a particular item than for instance with an alternative choice item and it is relatively straightforward to convert the responses into numerical form (consequently it is amenable to statistical analysis). Using the features I had obtained previously, I generated list of attitude items, for example it requires no prior knowledge of accounting became "I didn't need any knowledge of accounting to use the system". The Likert Scale, a particularly common type of rating scale, was chosen. It consists of a single continuum for each item with five positions running from *strongly agree*, *agree*, *uncertain*, *disagree* to *strongly disagree*. It was preferred because it allows respondents to express their level of agreement or disagreement with each item on an ordinal scale and therefore ensure that all items would measure the same thing. Other reasons for its selection are that it is certainly less laborious than other rating scales and was also used successfully by Adelman (1992) in his evaluation questionnaire.

8.3.2 Collection of demographic and background information

The other important part of the questionnaire was to obtain relevant demographic information from the respondents since it helps with the analysis at a later stage. The major reason is that it enables one to find out whether the group was representative as a whole and if individuals showed extreme or rare responses due to other factors (apart from the independent variables that allowed to vary). Accordingly space was included for the respondent to fill in name, age, sex and other background information.

8.3.3 Piloting the questionnaire

Once the first draft of the questionnaire had been produced, every item was carefully examined for grammar and clarity. The selection and editing of questions was governed, on the one hand, by attempting to collect data with clarity and the need to maintain a sufficiently high-level of interest in the questionnaire, consequently there were technical trade-offs between these two judgments. For instance, some questions were shortened,

because questions containing more than one idea should be avoided. In addition, I used the guidelines provided by Kidder (1981), Kerlinger (1973) and Oppenheim (1986) when scrutinising the items to prevent common errors such as ambiguity, imprecision or presumption. An example of such an error might be a leading question where the following is asked, "Do you not agree that parents should have the right to have a say in their children's education?" (i.e. most, if not all, respondents would answer yes).

My first task was to print out a draft form of the whole questionnaire (including questions about the background of the student) and try it out on an actual student(s). The purpose of the pilot exercise was to get any mistakes out of the instrument. I piloted the questionnaire to check several things, firstly how long it took recipients to complete it, to check all questions and instructions were clear (i.e. removed or edited questions which did not yield any usable data). Ideally, the respondents for the pilot should be similar to the population of the study. I obtained helpful feedback after a one-student pilot (Luke Howells, an undergraduate who had recently taken an accountancy module) organised by my supervisor. This exercise elicited very useful comments from the student not only about the questionnaire but also the system after he had used it. A large number of improvements were made to the questionnaire, for example the item "At least with a book you can flip backwards and forwards." was replaced with "I prefer an accountancy book because you can flip it backwards and forwards." I was also able to improve the questionnaire by asking for a critical review from two people with experience of questionnaire design within the department, namely Angela Sasse and David Fulton. I requested them to answer the following questions about the questionnaire (adapted from Bell, 1987):-

(1) How long did it take you to complete?

(2) Were the **instructions** clear?

(3) Were any of the **questions** unclear or ambiguous?

If so, will you say which and why?

(4) Did you object to answering any of the questions?

(5) In your opinion, has any major topic been omitted?

(6) Was the layout of the questionnaire clear and attractive?

(7) Do you have any further comments?

It allowed me to make significant revisions of the questionnaire and make clear preparations for the first large study.

8.3.4. Improvement and refinement of questionnaire appearance and layout

Ideally a good standard of questionnaire should present little or no difficulty to complete, look attractive, and manage to capture the required data. Therefore I took a considerable amount of time not only getting the content but also the presentation correct. I took critical advice from others, but also I scrutinised other people's questionnaires for things to do and for things not to do. Here is an indication of some of the guidelines I followed, others are given in Bell (1987) and Oppenheim (1986):-

Firstly, instructions should be clear (e.g. using either uppercase or a different typeface to the rest of the text) and you should remember that what is easy to understand for you may not be the same for others. Secondly, the care ought to be taken with the order of the questions, making sure easy ones appear first and leave the more complex and sensitive ones to the end. Thirdly, response boxes should be kept towards the right hand side and dotted lines provided for more detailed responses. This enables the respondent to reply more easily, at the same time helping the researcher to extract the data at the end of the study. These directions meant that I chose Helvetica font for the instructions since it made them easier and clear to read, whereas I chose Times Roman font for the actual questions because the reader must concentrate on individual wording. In addition, the section headings were in bold typeface to emphasise a new set of questions. Similarly I decided

to ask the simple demographic questions at the beginning and left the personal judgements till last. Furthermore places to respond were kept on the right-hand side of the page and I numbered the boxes with the appropriate question numbers.

Guidelines such as these were particularly important since my questionnaire was quite long and I had to make it appealing so that the respondent's interest was sufficient to keep going. I managed to enhance the questionnaire further by experimenting with what I had, in terms of font size, spacing and rearranging to see what looked best. As a result of this I was able to create a good working copy of the questionnaire.

The next step was to refine the questionnaire further by getting comments from the original developers of the software. So just before the Summer term I took my working copy and went through it in detail with the team working in Leicester, to check the wording and the format. This also proved useful as a number of technical errors were identified and corrected. For example I was able to rephrase the last question on the questionnaire from "If there was **one** improvement you could recommend...." to "If there was **ONE** outstanding improvement you could recommend....". Finally the questionnaire was redrafted several times and prepared for a pilot study. At the end of a variety of versions and drafts I had reduced the questionnaire down to 107 clear and concise questions about the system, from an original total of 132 rough statements (see appendix A for the version used in the study).

8.4 Investigating assessment tests for marginal costing

My objective was to develop a robust and feasible assessment test for marginal costing. My design dictated that there should be three tests, one pre-test which was relatively easy to solve, which would be given to all the participants prior to their exposure to different teaching sessions, but also two post-tests, one for the experimental group and the other for the control group. Two post-tests were chosen to avoid the possibility of cheating, since one group would receive the post-test before the other. Both these post-tests would be based on the same format but use different values and be slightly more difficult than the

pre-test. The intention would be that all the questions in the tests would be broken down into smaller questions to allow the less able to show their ability whilst also allowing the most adept at marginal costing to stretch themselves. In common with most examinations I did not want people to complete the test early so they would be fidgeting about for the rest of the time. So the tests fulfilled two major roles: (1) to take into account individual differences and (2) the measurement of improvement. The former was accounted for by the pre-test given to all participants and the latter was accounted for by the post-tests. *The choice of a different post-test from the pre-test was confirmed as a rational and correct choice by Professor Mel Slater (UCL research tutor) and significantly by Dr Tom Fearn (UCL Statistics Department) whose speciality is the design of experiments. In fact, Dr Fearn was kind enough to supply a reference from the published literature establishing this beyond doubt i.e. page 55 in Cox's (1958) book "Planning of Experiments".*

My initial plan was to assess the students performance by giving them a pencil and paper test, using past examination questions on marginal costing, such as the ones given to undergraduates at De Montfort University. However while talking about the questionnaire to Angela Sasse, I felt that this might not have been enough. Subsequently I needed to re-examine the appropriateness of my approach. I carried out some background reading and tried to find out about some of the various ways to assess performance. Therefore I read material such as Frith et al. (1984), Ebel et al. (1991) and Tyler (1973) in order to get a good grasp of what alternatives were available. A lot of the literature was confusing since different theories of education propounded their own style of tests. I approached De Montfort University for a set of questions. However I found out after discussing with another lecturer that these questions could be solved in other ways apart from using marginal costing knowledge. Consequently I decided to settle on a set of questions adapted by Julian Lavis (Thames-Valley University) from past accountancy examination questions given at the end of the chapters in Lucey (1992).

8.5 Contacting and arranging meetings with accountancy departments

The most difficult research problem facing me, apart from obtaining *fully-functional* teaching software, was to get hold of a large number of students to test the system. This would mean obtaining the co-operation of a number of educational institutions. Since University College London (UCL) did not teach accountancy (I found out later through my supervisor that a new modular course included accountancy and particularly marginal costing), I had to telephone and write to other universities in the local area. Originally I tried London School of Economics and also University of Westminster. Making contacts with these took a long time, so I broadened the search. Some places were helpful, like Thames-Valley University, while others were surprisingly quite hostile. My main procedure has been to make contact with the person in charge of the undergraduate course and then meet them to explain my research in more detail and demonstrate the software. This approach lead to positive responses, since the person is able to try out the system for themselves and also ask any questions relating to the research. This process continued, with follow up meetings and chasing up the non-respondents. In addition, once I was given the go-ahead was given a lot of issues needed to be confirmed such as size of experimental and control group, timing of assessment (prior to teaching marginal costing topic), and availability of computer labs. It was an extraordinary task since I relied on the commitment of lecturers to provide valuable time in their teaching schedule, book classes and gain access to their computer networks. I have to say that I did come to have deep appreciation for those that did help.

Since the evaluation also needed the active cooperation of accountancy students and they would be required to contribute their time and effort, particularly difficult for students in a busy academic timetable and with a heavy workload. It was considered that for them to just know they have contributed to the search for knowledge or learning something about themselves would not be enough to encourage their participation. Especially given that people have the right to know what they are getting into and the right to give or withhold their cooperation on the basis of that information. Consequently, it was reasoned that in order to increase the benefit to them, remuneration from the graduate school fund would be sought. Normally in human research, people are often paid for their participation. However by the time I found that the issue of funding needed to be addressed, there seemed to be a complete lack of knowledge about its source from my department and this

delay in acquiring the information was to prove influential (as we shall see this had an effect on the number of people who took part).

In order to prepare for the study each university lecturer with a positive response was sent a letter outlining the details of the proposed study and a short questionnaire asking for details of the students e.g. number of likely participants and their course (to be returned after completion) so that the logistics of the study could be worked out. In addition a meeting was arranged after reception of this questionnaire to discuss how the study could be carried out. At this final meeting I would complete negotiations and plan logistics with the help of university lecturers, such as which rooms would be available to carry out the study on the experimental group as well as the follow-up for the control group. The university lecturer had of course to be in a computer lab since I needed access to personal computers to run the marginal costing package. In addition, I needed to know information on number and names of UCL students who were able to participate. For one site, I had to provide a letter from the developers of the software for permission to run the software. One particular problem arose quite soon after I had made the arrangements for the study to be carried out at the first site. This was that the copy I had was meant to run on stand-alone personal computers, however since it was originally designed using DOS version 4.X and the networked computers I intended to run this software used up this area of memory therefore it had refused to run on them. I was running out of time, and I had to carry out the study according to the teaching timetable, otherwise my preparations would be wasted. Luckily I managed to get in contact with systems support at the ISD in UCL and they were able to adapt the software so I could run it. However instead of installing it on to the hard disk, I had to run from the diskette. I found out later this was to have an effect on the speed of use.

For each study I needed to copy two versions of the questionnaire (one for the control and experimental group), three different tests (pre-tests, two post-tests) and take enough diskettes of the marginal costing package. In fact I had to make sure I had more than enough copies to go around. Also I took a whole bag full of other items: spare paper, pencils, guidance notes for myself, two calculators, chalk, a white board pen, a sign for

the door (saying "Please Do NOT Enter!") attached with *blu tack* (the tradename for a reusable adhesive).

8.6 Carrying out the study at each site

In the study I started off by introducing myself, explaining my research and describing what I intended to do in the next hour or so. Both experimental and control groups were present (I checked off the attendance from a list of students provided by lecturer at an earlier meeting). I described the overall aim of the work was the assessment of a computer-aided learning program. Then I emphasised the results from this study would have no effect whatsoever on their course and that all information provided by participants would be confidential. If anyone was interested in further information they could contact me at computer science department of University College London. Then I went on to describe the events that would take place in the next half an hour. They would be given a detailed questionnaire to complete and two questions on marginal costing (based on a textbook question) to answer. Then I asked if there were any questions.

I handed out the pre-test in the style of an examination, that is face down, and asked them if they required pencils. Before I asked them to start, I explained they had 10 minutes to answer this question fully. They may use calculators but they were required to show each stage of their working out clearly and to make sure they read each page carefully. At the end of this period, all the questions were collected by me. Then the group was split into two (using alphabetical selection from the student list) - the experimental group remained behind and the control group left.

The experimental group was given an introduction to the software by gathering around a personal computer. This enabled me to demonstrate all the important functions of the system, that is the tutorials, the exercise (or problem generating) component as well as the calculator and graph facilities. They were given 30 minutes to familiarise themselves with ALL these aspects. In addition I went around and made sure each student had at least read

through two basic tutorials, solved two problems (questions) and used the calculator and graph facility.

Following this I gave the experimental group an introduction to the questionnaire, carefully explaining what type of information I was seeking in each of four sections, for example in section D I wanted them to provide their frank and critical assessment of the software. In addition, I explained the way to use a Likert Scale, for those unfamiliar with it. This I did by giving them an example on the white-board. A Likert Scale was drawn on the board and below this was a simple statement "Mickey Mouse is an excellent cartoon". Then I explained that if I *disagreed* with this statement I would respond with a *4* in the box (next to the statement). Similarly if I *strongly agreed* that this statement was true then I would answer *1*. However, I did *not* want them to put any values in between the ones given on the scale such as 1.5 or 4.7 (i.e. only whole numbers). They were given 20 minutes to complete the questionnaire. On a few occasions there were some foreign students so I was able to answer any queries about question wording.

Finally they were given the post-test, which included slightly harder questions in order to check whether they had improved on their marginal costing knowledge. This again lasted for 10 minutes. At the end I thanked them for their time and help and handed out a contact address and telephone number should they want their results or more information about the study.

The control group were contacted again following their conventionally taught period of marginal costing. They were given a smaller version of the questionnaire (i.e. without the section on the assessment of the software) and asked to complete this within 10 minutes. After which they were provided with slightly amended version of the post-test (same question but different values) and asked to complete this within 10 minutes. Again they were thanked for their participation and details provided if they wanted to get in touch for further information.

CHAPTER NINE

RESULTS

"A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it." [Max Planck *Scientific Autobiography and Other Papers*, translated by Frank Gaynor, p33-34, 1950] (18;18;5;20;21;18;14;20)

9.1 Introduction

The purpose of this research was to assess the value and effectiveness of computer-aided instruction (CAI) systems with target users. This investigation was conducted in the hope it will provide CAI designers with the required information to improve on their products. Two aims of the research emerged from this overall goal. The first aim was to find out whether the system was at least as effective as conventional teaching. The second aim was to determine which key features or characteristics (i.e. criteria) of the system are significant in assessing the quality of a teaching system (where quality is defined as *fitness for purpose*). The assumptions underlying this research are that conventional teaching is a suitable comparative standard to aim for and that we can obtain useful criteria from the users critical assessment of the system.

The practical part of this work was carried out using a fully operational accounting tutor (which teaches marginal costing). This system was developed in De Montfort University and is currently being used in a number of universities as part of their normal teaching programme for accountancy students. This study was carried out with intended users of the CAI system, i.e. undergraduate and postgraduate accountancy students. Students participating in the study came from three different universities: University College

London, University of Westminster and Thames Valley University. They were all studying a course with a component of Marginal (or Variable) Costing. At each site, students were split up into two groups. One group was the *experimental* group which was allowed to interact with the marginal costing software, while the other group acted as the *control* and therefore not allowed access to the software. All the results from both the questionnaire and tests were coded and entered into the Statistical Package for the Social Sciences (SPSS) for analysis.

9.2 Descriptive Statistics

In total, forty-two students participated in the research, twenty-one from University College London, twelve from Thames Valley University and nine from University of Westminster.

9.2.1 Description of the Participants

A profile of the respondents was built up by running two main sets of statistics on the data collected from the first three sections of the questionnaire (i.e. biographical data, current and previous experience with computing). Firstly a number of basic statistics (i.e. mean, standard deviation, minimum and maximum) were obtained and secondly a simple frequency command was run in order to find how many people gave an alternative response to each question (frequency distributions for all the questions in the questionnaire are given in appendix C).

Apart from a minority of postgraduates (i.e. from the University of Westminster), the *greater part* of the students (78.6%) were *undergraduates*. *Approximately half* (54.8%) were male participants, in fact 19 out of 42 of the participants were female. The ages of the participants ranged from 18 to 35 years. The frequency distribution for age (given in appendix J) shows a positively skewed bimodal distribution, with *peaks at 19 and 21*

years. The majority of students (64.3%) were under 22 and the most typical age was 21.9 years old.

The order of birth in a family was particularly interesting because it has been found to be related to high achievement. People who are first-born in a family are *likely* to be high achievers, indeed the majority of US Astronauts were found to be first-born. Consequently, this question was chosen to see if this variable had an effect on performance or assessment of the system. The results showed that the order of birth (given in appendix J) ranged between being born first and sixth in the family and out of this 61.9% of participants were either first or second-born. Significantly, *more than half were first-born* and therefore were likely to be high achievers. An unsurprising fact given that all participants had managed to gain admission to university.

A number of questions scrutinised whether the lack of computer experience had any impact on the students' performance or assessment with the CAI. This was important since this particular system was conceived as something which required minimal computing knowledge - if any. Over half of the participants (52.4%) expressed a clear preference for enjoying using the computer *quite a lot*. Nearly everyone (except one participant) acknowledged to *having access* to a computer at college or work. The majority of these respondents revealed they could use the computer (90.2%) or be able to operate it by themselves (70.7%). However less than two-thirds (59.5%) of participants had a computer at home and about 80% of these admitted to using it or knowing how to operate it without assistance. Clearly most, if not all, accountancy students had access to computers and a working knowledge to use them but they *could not be considered as keen computer users*.

The participant's reluctance to use computers seems to have been *sown at an early age*. When a question was raised about their experience of computers at school very few (2.4%) had used it daily, 33.3% had used on a weekly basis, 11.9% had used it on a monthly basis, 33.3% had used it rarely and 19.0% never had contact with it at all. As for specific experience with computers following school, a large number (69%) had received *some form of computer education or training* and the most popular forms of software packages were spreadsheets and wordprocessing. Moreover, the students could *not be*

considered ardent programmers, since only 26.2% could program in one or more computer languages and typically the greatest number (11.9%) admitted to only being able to program in one language.

9.2.2 Participant's assessment of the Marginal Costing System

Using the frequency distributions, the attitude ratings for the CAI system were displayed as percentages. On the whole students gave a positive assessment of the marginal costing system. They were, however, critical about certain aspects of the system and indicated they were unsure about other features. A detailed examination of the results is presented on the following pages by focusing on the responses given to a number of related statements (grouped under the same headings as the questionnaire).

COMPUTING KNOWLEDGE - A major claim made by the developers was that this system does not require any previous knowledge of computing. Respondents were asked whether computing knowledge was an obstacle in using this system or if this was taken care of by the design. Their overall attitude endorsed the view of the developers.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
14. This program does not require any prior knowledge of computing.					
14.	8.7%	60.9%	21.7%	8.7%	
15. I didn't learn anything because I was spending all my time getting to know how to use a computer.					
15.	4.3%			52.2%	43.5%
16. After using this system I became more confident about learning from a computer.					
16.	4.3%	47.8%	30.4%	13.0%	4.3%
17. It was difficult enough to tackle the accountancy problems without having to fight with the computer.					
17.		8.7%	4.3%	56.5%	30.4%
18. The system can be used sufficiently well without any manuals.					
18.	8.7%	60.9%	21.7%	4.3%	4.3%
19. Normally I find this type of subject very simple, but I couldn't see what the computer wanted me to do.					
19.	4.3%	8.7%	4.3%	56.5%	26.1%

- When asked, in three different ways, whether the computer presented them with problems in learning accountancy they vehemently disagreed every time.
- They agreed that prior knowledge of computing was unnecessary and the system could be used without manuals.
- They showed less certainty, however, when asked if they were "confident about learning from a computer" following contact with this particular system, since only a small majority agreed with this statement and many others did not share this conclusion.

COMPARISON WITH OTHER MEDIA - A group of statements were used to find out whether the respondents found learning from this system more difficult in comparison with other forms of media, such as one-to-one teaching, peer consultation and textbook study. On the whole, respondents compared the system favourably since it was more interactive and flexible than a book but comparisons with lecturers were divided.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
20. This system is a better teacher than most lecturers.					
20.	13.0%	17.4%	39.1%	21.7%	8.7%
21. Having a lecturer close by, so that I could ask any questions, would have been a big improvement.					
21.	21.7%	34.8%	26.1%	17.4%	
22. I liked understanding marginal costing with this system, because it's like being on a one-to-one basis with your tutor.					
22.	4.3%	56.5%	26.1%	8.7%	4.3%
23. I was left with a lot of unanswered questions after using the system.					
23.		13.0%	26.1%	43.5%	17.4%
24. This form of teaching was really clear, because unlike a lecturer teaching, the program does not miss out a lot of steps.					
24.	13.0%	21.7%	47.8%	13.0%	4.3%
25. It is easier to get involved solving accounting problems using this system than in a classroom tutorial.					
25.	21.7%	39.1%	13.0%	21.7%	4.3%
26. I prefer an accountancy book because you can flip it backwards and forwards.					
26.		13.0%	21.7%	56.5%	8.7%
27. It's much easier to understand than reading an accountancy book because it's interactive.					
27.	34.8%	34.8%	26.1%	4.3%	
28. It was terrific to be able to discover the relationships between variables, which I never got from books.					
28.	17.4%	47.8%	17.4%	17.4%	
29. I would rather learn accountancy by working out problems using pencil and paper.					
29.		8.7%	39.1%	39.1%	13.0%
30. I like doing accountancy this way because you can see how it's done, rather than having to ask somebody to explain it.					
30.	8.7%	56.5%	13.0%	17.4%	4.3%

- Respondents rated this system moderately better than accountancy textbooks, they found relationships between variables were more apparent and it was more interactive than a book. They disagreed that books were desirable because they could be easily manipulated. They also disagreed that accountancy problems are best done by hand, although this response was not so straightforward.

- Comparisons with lecturers and tutorials were less clear-cut. Respondents were divided on whether having a lecturer close by when using this system would have been an advantage, or if this system is better since it does not miss out teaching steps, or even if overall it is a better teacher than most lecturers. Where there was partial agreement, was that the system resembled one-to-one tutoring. There was also a small recognition that this system was easier to get involved with compared to classroom tutorials.

- Moreover this system was seen as a favourable way of doing accountancy compared to asking someone to explain it and most respondents were inclined to disagree they had any unanswered questions following use of the system.

INDIVIDUAL FLEXIBILITY - Several statements examined how well the system was able to cope with the diversity of user demands. There was a mixed response.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
31. This system is only appropriate for advanced students of accounting.					
31.		4.3%	17.4%	60.9%	17.4%
32. The computer got harder just as I got better at the problems.					
32.		30.4%	30.4%	39.1%	
33. I understood what I had to do, but the system persisted in following its long and laborious steps.					
33.		4.3%	34.8%	60.9%	
34. The system acts like a human tutor because it adjusts its advice according to the individual's needs and progress.					
34.	4.3%	39.1%	21.7%	26.1%	8.7%
35. I found it helpful to jump into the more demanding material without having to go through the easier stuff.					
35.	8.7%	56.5%	30.4%	4.3%	

- The majority of respondents agreed that this system matched their own ability level, by accepting it did not follow long and laborious steps and it was not just for advanced respondents. They also moderately agreed that it enabled them to go directly to the more difficult material.

- Very few, however, were convinced that this system was adaptable enough to adjust its advice like a human tutor or that it stretched more capable users with more difficult problems.

ACCOUNTANCY KNOWLEDGE/TEACHING - A key aspect of this system is how well marginal costing is catered for, including whether prior accounting knowledge is essential to its use and whether the topic is adequately structured for the learner. Most found the system was able to give a basic understanding.

- There was strong agreement the system provided a first class grounding in the use of marginal costing information and guidance through the learning process. It was also strongly acknowledged that the respondents were able to fully test their understanding of marginal costing and gain an accurate picture of marginal costing. Unsurprisingly therefore, there was clear disagreement that learning marginally costing basics from this system was difficult or that on completion they still did not know anything about marginal costing.

- The respondents were moderately satisfied that this system gave a good introduction to marginal costing concepts, it encouraged them to consider the relationships behind marginal costing, their knowledge of accounting was greatly increased and they would like to learn other accounting topics using this system. As to whether the program is only useful for people who are already familiar with accountancy, or that they were left without any understanding of the relationships between different parts of marginal costing, or that this system does not allow most people to work out marginal costing problems, there was only modest disagreement.

- It is worth noting less confidence was shown whether this system left them enough knowledge to work out marginal costing, using real information, or even to teach someone else about the subject using this system.

- The majority were uncertain whether users could become experts simply by using this system and an ambiguous response was also recorded when asked if no previous knowledge of accounting was needed to use the system.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
36. I didn't need any previous knowledge of accounting to use the system.					
36.	17.4%	34.8%	17.4%	26.1%	4.3%
37. I have gained a good introduction to the concepts of marginal costing.					
37.	13.0%	60.9%	21.7%		4.3%
38. I think the program is only useful for people who are familiar with accountancy and want to improve.					
38.	8.7%	13.0%	26.1%	43.5%	8.7%
39. At the beginning I was not sure what marginal costing meant, but once I got the hang of it, I became quite an expert.					
39.		17.4%	56.5%	17.4%	8.7%
40. It is difficult to learn the basics of marginal costing with this system.					
40.			8.7%	87.0%	4.3%
41. The system gives a first class grounding in the use of marginal costing information.					
41.	8.7%	60.9%	30.4%		
42. Most people would find it hard to work out marginal costing problems using this system.					
42.		8.7%	26.1%	52.2%	13.0%
43. The system allows you to fully test your understanding of marginal costing.					
43.		73.9%	21.7%	4.3%	
44. At the end I still did not know anything about marginal costing.					
44.		8.7%	4.3%	43.5%	43.5%
45. My knowledge of accounting has been greatly increased by using such a system.					
45.	4.3%	47.8%	21.7%	21.7%	4.3%
46. After using this system I can easily use my knowledge to work out marginal costing for most information given about a firm.					
46.	4.3%	43.5%	39.1%	4.3%	8.7%
47. I would find it easy to use this system to teach someone else about marginal costing.					
47.	8.7%	47.8%	34.8%	4.3%	4.3%
48. The computer has helped me by guiding me through the learning.					
48.	4.3%	82.6%	4.3%	8.7%	
49. I would like to learn other accounting topics using this system.					
49.	34.8%	34.8%	21.7%		8.7%
50. The computer system allowed me to learn an accurate picture of marginal costing and correct my mistakes.					
50.	17.4%	73.9%	4.3%	4.3%	
51. It allowed me to think about the relationships behind marginal costing.					
51.	8.7%	56.5%	26.1%	8.7%	
52. Even now, I do not understand the relationships between the different parts of marginal costing.					
52.	4.3%		26.1%	43.5%	26.1%

ERROR MESSAGES AND DOCUMENTATION - The significance of error messages are that they guide the user, e.g. from making inadvertent slips where the user lacks knowledge about or misunderstands the aspect of the system being worked on.

- Respondents overwhelmingly agreed the messages helped them when they got stuck and told them what they did wrong and even gave advice after allowing them to make mistakes.

- They moderately disagreed the program was unforgiving and that incorrect values should be excluded. They agreed it told them precisely what they did wrong and an error message for the incorrect use of a positive or negative number should be included.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
53. The computer messages were good enough to help me when I got stuck.					
53.	17.4%	69.6%	8.7%	4.3%	
54. The error messages were unhelpful.					
54.		4.3%	17.4%	52.2%	26.1%
55. I liked the program because it allows me to make mistakes, so that it can advise me what to do.					
55.	39.1%	52.2%	8.7%		
56. The program should not allow you to enter any incorrect values.					
56.	8.7%	8.7%	13.0%	34.8%	34.8%
57. The program is so unforgiving, it doesn't allow you to do anything else, till you provide the correct answer.					
57.	8.7%	8.7%	21.7%	43.5%	17.4%
58. There should be an error message if I enter an incorrect negative or positive numbers.					
58.	17.4%	47.8%	21.7%	8.7%	4.3%
59. I like the way it told me precisely what I did wrong.					
59.	26.1%	52.2%	13.0%	8.7%	

EXPLANATION (FEEDBACK) INSTRUCTIONS - Users need informative feedback and the provision of appropriate explanations is a significant part of this communication process. For the majority of occasions they were pleased with the content of feedback.

- Respondents considerably favoured the feedback because they immediately knew if the answer was correct and rejected the assertion that explanations were hard to understand. They were reasonably satisfied that the system helped them with every step in solving the problem.

- There was partial agreement that good feedback was only provided once an answer was entered and criticised the suggestion there was not enough praise for correct answers.

- Surprisingly though, most respondents were uncertain whether the explanations were clear enough to show why the answer was wrong or if the program gave individual feedback to each respondent or if a clearer explanation of the aim of the exercise and more information would have been better.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
60. The system only provides good feedback once you have entered an answer.					
60.	8.7%	47.8%	34.8%	8.7%	
61. The system didn't give enough praise for the correct answer.					
61.	4.3%	21.7%	26.1%	30.4%	17.4%
62. The system helped me with every step I took in solving the problem.					
62.	4.3%	52.2%	26.1%	17.4%	
63. Through the feedback I immediately knew whether I had got it right or wrong.					
63.	30.4%	60.9%	4.3%	4.3%	
64. The explanations given by the program are hard to understand.					
64.		4.3%	21.7%	69.6%	4.3%
65. The explanations showed clearly why I had got the wrong answer.					
65.	8.7%	30.4%	39.1%	21.7%	
66. The program gives individual feedback to each student depending upon how far they have progressed to the solution.					
66.	4.3%	30.4%	52.2%	13.0%	
67. A clearer explanation of the aim of the exercise and the information needed to complete it would have been much better.					
67.	8.7%	21.7%	43.5%	26.1%	

ROUTE THROUGH PROGRAM - One indication of how well users were comfortable using the computer, as well as using it to apply their accountancy problem-solving skills, was to ask respondents how easy it was to negotiate through this learning environment. Most were satisfied with the system but a sizeable number preferred more direction.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
68. I never knew what all the items on the menu meant.					
68.	4.3%	17.4%	17.4%	52.2%	8.7%
69. The instructions given on how to proceed through a tutorial were unclear.					
69.	4.3%	8.7%	26.1%	52.2%	8.7%
70. I like learning like this, because you can check over things and not make the same mistakes.					
70.	13.0%	52.2%	30.4%	4.3%	
71. Sometimes I found it hard to keep track which bits I had done.					
71.	4.3%	13.0%	17.4%	60.9%	4.3%
72. It allowed me to solve a problem in any order.					
72.	26.1%	43.5%	13.0%	17.4%	
73. The system is unable to recognise a different route to the solution.					
73.	8.7%	13.0%	52.2%	26.1%	
74. I would have found it more helpful to be given a suggested route through the program.					
74.	8.7%	39.1%	30.4%	21.7%	
75. The prompts, asking me to enter text, weren't clear from the screen.					
75.		8.7%	21.7%	60.7%	8.7%
76. I want to be able to stop the exercise and go back to the tutorial, and then return where I left off.					
76.	30.4%	47.8%	17.4%	4.3%	
77. On-screen tutorials are difficult to follow.					
77.	8.7%	8.7%	17.4%	65.2%	
78. It provided little understanding of what you are doing or how to do it.					
78.		26.1%	21.7%	52.2%	

- On the majority of occasions they had no major problems, therefore they disagreed that tutorials were difficult to follow, instructions through a tutorial were unclear, or that it was hard to keep track, and prompts were vague. A sizeable number, however, did report they did need a refresher facility half way through the exercises.

- Slightly ambiguous responses were given when asked whether the system could cope with alternative paths to problem solving, i.e. "system is unable to recognize a different route" and "it allowed me to solve a problem in any order".

- A moderate number of respondents agreed this form of learning reduced the chances of making same mistakes, however, others would have preferred a suggested route through the program.

CONTROL OF LEARNING - A few statements were developed to find out whether the learner could manage the system.

- A large consensus agreed that the system provided the correct pace and direction of learning and the majority did not find it too slow or too quick.
- The question about "trying different things with the system" did generate some confusion.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
79. I felt the system allowed me to work at my own pace and direction.					
79.	26.1%	73.9%			
80. I found the system was too slow because I could do the exercises much more quickly with pencil and paper.					
80.			21.7%	52.2%	26.1%
81. It was too quick, I would like more control of how much time I was given to solve the problems.					
81.	4.3%	4.3%	4.3%	69.6%	17.4%
82. I learnt a lot about marginal costing by trying different things with the system.					
82.		47.8%	43.5%	8.7%	

RETENTION OF LEARNING - Respondents were asked to own assess how much useful knowledge they had gained through using the system.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
83. I am sure I will retain most of what I have learnt.					
83.		43.5%	39.1%	17.4%	
84. I learnt most from this system, by trying to answer the various problems.					
84.	21.7%	60.9%	13.0%	4.3%	
85. If I use this program again, I could solve similar problems much quicker.					
85.	21.7%	56.5%	21.7%		

- When asked particularly whether they would solve problems better following exposure to this system an overwhelming number agreed, however, when asked whether their general level of knowledge had been increased the response was less certain.

HELP FACILITY - Apart from error messages, the help facility is the only way a system designer can ensure users do not misunderstand what they should be doing.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
86. I would prefer more help from the system.					
86.	26.1%	13.0%	39.1%	21.7%	
87. The help facility should have given a few hints, rather than the correct answer.					
87.	4.3%	39.1%	30.4%	26.1%	

- No pattern emerged from the results except that satisfied users and unsatisfied users may be disguised by the overall response and is therefore worth investigating further.

FUNCTIONS OR KEYS - Functions and keys need to be grasped easily and quickly, not only for users who are not computer literate, but also because the keyboard layout will vary with the computer used.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
88. I found it very difficult to know which keys corresponded to the command I wanted.					
88.	4.3%	13.0%	17.4%	47.8%	17.4%
89. When proceeding through the program I found I was clear which options I was allowed to select.					
89.	8.7%	69.6%	8.7%	8.7%	4.3%
90. I hated some of the option keys, I sometimes got lost because of them.					
90.		13.0%	17.4%	47.8%	21.7%
91. When I was going through the program, pressing the same key would produce a different command.					
91.		8.7%	34.8%	47.8%	8.7%
92. It was easy to exit from a particular route through the program.					
92.	21.7%	60.9%	4.3%	13.0%	
93. I couldn't skip some parts of the exercise.					
93.		26.1%	26.1%	43.5%	4.3%
94. It was simple to delete any answer which I knew was wrong.					
94.	26.1%	56.5%	4.3%	8.7%	4.3%
95. I wasn't allowed to restart with the same data.					
95.	8.7%	8.7%	60.9%	13.0%	8.7%

- Respondents confidently knew how to delete incorrect answers, as well as the options open to them and how to escape from the program at any stage. Moreover they disputed getting lost because of the option keys, or not knowing which key was appropriate, or being unable to skip some parts of the exercise, although less confidently than the previous statements.

- A majority disagreed (and a small number were uncertain) whether the keys were consistent throughout the program. An indistinct response was recorded when asked if it was frustrating at not being able to return to the exercise with the same data.

CALCULATOR FACILITY - A major requirement of any accountancy package is a calculation facility, in this system the computation of figures with ease was even more important since the system was designed to be used by people without accountancy knowledge.

- Responses about the calculator were the most clear-cut set of all the results obtained. An overwhelming majority found the calculator easy to use.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
96. The calculator facility was easy to use.					
96.	39.1%	47.8%	4.3%	8.7%	
97. I found it hard to use the calculator to transfer the answer to the program.					
97.		4.3%	4.3%	56.5%	34.8%
98. Using the calculator is frustrating - it doesn't do what it's supposed to do.					
98.		4.3%	8.7%	39.1%	47.8%

GRAPHICS FACILITY - There are two main functions of the graphics facility that is, to describe the relationship between the variables used in marginal costing and to give an on-going representation of the exercise being tackled. The system did this admirably.

- Respondents emphatically declared that it was not difficult to read information from the graph and in fact they could comfortably use it to compare the effect of different decisions.
- The respondents agreed that the graph was very useful, although this was less pronounced.
- When asked if more graphics, colours and sounds should have been included the answer was uncertain.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
99. I found the graph very useful.					
99.	17.4%	43.5%	30.4%	4.3%	4.3%
100. It was difficult to read the information from the graph.					
100.	4.3%		26.1%	60.9%	8.7%
101. The graphical facility allows you to compare easily the effect of different decisions on costs.					
101.	17.4%	56.5%	26.1%		
102. I would have enjoyed it better if there had been more graphics, colours and sounds.					
102.	17.4%	17.4%	34.8%	30.4%	

PRESENTATION INFO. ON SCREEN - The interactive screen was meant to compensate for a lack of computer skills. The overall reaction confirmed that the screen was well-designed for this task.

- There was definite acknowledgement that the screen layout made it easy to work out solutions to exercises as well as communicate with the computer.
- Respondents strongly disputed either that the colours were confusing or that they were confused by the amount of information on the screen.
- A consistent but less emphatic response confirmed the screen was attractive and clear, as well as dismissing there was insufficient information to solve the exercises.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
103. The screen layout meant that it was easy to communicate with the computer.					
103.	8.7%	78.3%	13.0%		
104. Sometimes I was confused by the large amount of information on the screen.					
104.	4.3%	8.7%	26.1%	56.5%	4.3%
105. The way the problem is presented on the screen makes it easy to work out the solution.					
105.	13.0%	60.9%	8.7%	17.4%	
106. There was not enough information on the screen for me to solve the problems.					
106.	8.7%	4.3%	30.4%	43.5%	13.0%
107. The screen was attractive and clear.					
107.	8.7%	43.5%	30.4%	17.4%	
108. The colours on the screen were very confusing.					
108.		4.3%	17.4%	65.2%	13.0%

HOLISTIC SYSTEM - A number of statements were used to gauge the personal impact of the system and their likelihood of using it again.

- Respondents strongly declared the system became easier with use and was a good way to revise. They rejected that the program went on too long or that it could not teach anything. There was, however, an ambivalent response to whether this computer program was better than any used previously.

- Respondents were happy to be given assignments using this program and found it helped them when in trouble. However, they also agreed that it could be improved a lot more. They liked using this system since it was fun to use and denied it was simply used in a non-interactive manner (i.e. to read off the screen).

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
109. I really liked using this system because it was fun to use.					
109.	8.7%	43.5%	34.8%	4.3%	8.7%
110. I enjoyed it but it could be improved a lot more, so I would be tempted to use it again.					
110.	13.0%	52.2%	21.7%	8.7%	4.3%
111. I have used a computer program like this before, but this one is a lot better.					
111.	4.3%	17.4%	52.2%	21.7%	4.3%
112. I simply used the computer to read what is on the screen.					
112.		17.4%	21.7%	52.2%	8.7%
113. The program went on too long.					
113.			21.7%	65.2%	13.0%
114. The more you use it, the easier it becomes.					
114.	17.4%	69.6%	8.7%	4.3%	
115. I would be happy, if I could be given assignments using this program.					
115.	13.0%	47.8%	21.7%	8.7%	8.7%
116. The program did not manage to convince me that I could learn something by using it.					
116.	4.3%		17.4%	60.9%	17.4%
117. I liked the way it helped me when I was in trouble.					
117.	13.0%	52.2%	26.1%	8.7%	
118. It is a good way to revise something I have forgotten.					
118.	21.7%	60.9%	17.4%		

OVERALL PERCEPTION - The last few items examined the general deficiencies of the system.

- Respondents strongly disagreed they had a poor understanding of the problems despite solving them. They agreed more examples would have been preferable. While a minority reported their colleagues would be interested in using such a system, even though most were equivocal.

Question	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
119. Most students are not interested in using such a system.					
119.			56.5%	34.8%	8.7%
120. I solved the problems but I was unclear why the answer was correct.					
120.	4.3%	4.3%	17.4%	65.2%	8.7%
121. I would like to have more examples.					
121.	8.7%	52.2%	17.4%	17.4%	4.3%

9.2.3 Responses to the Open Questions

The open-ended format, at the end of the questionnaire, was an opportunity for the respondents to suggest any features (except those already mentioned) which result in the strength or weakness of the present system and any single improvement they would make. A variety of responses were classified into similar groups and ranked in order of frequency. A full frequency table for each question is listed in appendix D and a selection of the most common responses is provided below.

RESPONSE	FREQUENCY
Quick & Easy to use, lack of computer knowledge does not cause problems.	5
Lots of good examples and exercises.	5
Clarity of topics, clear and logical, good program design, main points on individual topics.	4
Self-explanatory, told me answer was wrong & how to work it out.	3
Good pace of learning & step by step methods.	3
Simplicity.	2

Table 9.1: What features did you find contributed to the STRENGTH of this system?

On the whole the response was favourable, however, from the results we can clearly observe that the user interface was a very important factor for the system users. An inconsistent or poor user interface can be a significant indication of an inferior quality product. Positive attributes of this system were that it was *quick and easy to use*, *self-explanatory*, possessed a *clarity of topics* and *step by step methods*. Evidently the careful arrangement of lessons within the CAI system was seen an underlying reason for its success. Another important factor proved to be the number of examples and exercises provided. The system's attractiveness in this case may be originating from the capability of less able accountancy students to gain a good grounding in marginal costing through their involvement with the problems and exercises. Some evidence for this, appealing level

of usability, is provided by examining the full list of responses, particularly as the following advantages were mentioned: *involvement in problems/exercises, interactive and in short time learnt basics*. In summary, students were engaged by the usability, interactivity, exercises and examples of the system.

RESPONSE	FREQUENCY
More elaborate and self-explanatory instructions, not enough information to carry on, not user-friendly.	4
Error messages ambiguous.	3
Slow (due to floppy disk), requires too much memory.	3
Long texts/tutorials.	2
Poor feedback (more praise).	2
Problems set in a context, more incorporation examples into tutorials.	2
Should have print command, for basics like formulas.	2

Table 9.2: What features did you find contributed to the WEAKNESS of this system?

The main criticisms were: *inadequate instructions, ambiguous error messages and exercises that should have been incorporated into tutorials* (instead of separating them). These respondents seem to be inviting a more discernable approach which includes a better help mechanism and a more sophisticated arrangement of lessons. One artifact of the study was the problems reported with the speed of the computer, this was because it was only possible to run the program off floppy diskettes at the university sites. A few suggested a print out command for the tutorials which indicates they are using the system like a textbook.

Suggested improvements to the system to some extent reflected the user interface again but other types of deficiencies were also cited. Most respondents wanted more elaborate instructions (signifying poor error handling by the system and therefore poor quality) as well as exercises to be included with the tutorials. Other suggested included a more

attractive screen design and even a few pointed out they had not really been given enough time to use it to make a useful judgement. This leads us to conclude that a number of well-placed adjustments would undoubtedly lead to an even more favourable picture.

RESPONSE	FREQUENCY
Elaborate and explanatory instructions.	4
Make screen attractive, better graphics and colours, too rigid graphics.	3
More examples, keep previous exercises for review.	3
Examples in context, incorrect answer should lead to relevant information.	2
Cannot comment (due to lack of use).	2
Speed it up, run on hard disk.	2

Table 9.3: If there was ONE outstanding improvement you could recommend to this system, what would it be?

9.3 Inferential Statistics

9.3.1 Pearson's Correlation for inter-rater reliability

The examination-type questions, that were given to the students in the pre-test and post-test, were marked by the researcher and an independent person (denoted by A and K respectively). The marking scheme consisted of a point for the correct answer and another point if the student had managed to indicate they were on the right lines by their working out (e.g. if they had used a suitable formula). Since the pre-test was marked out of a total of 8 and the post-test was marked out of 10 (the post-test was longer and more involved than the pre-test), it was felt necessary to normalise the pre-test scores, that is as if they had been marked out of 10 to allow fair comparison (the normalised variable names started with an "N").

Typically, inter-rater reliability can be assessed by calculating the degree of association (correlation) or agreement (concordance) between the two sets of measurements. Pearson Product Moment Correlation (as opposed to Spearman's Rank Correlation) was chosen because the scores for each student were classified as interval data. Moreover, since we had hypothesised that the scores would be correlated we chose a one tail of significance. The pre-test scores and post-test scores were compared separately for inter-rater reliability (given in appendix E). The Pearson correlation for scores from the:-

Pre-test (using students $n = 42$) Post-test (using students $n = 42$)

Pearson $r = + 0.98$

Pearson $r = + 0.94$

Both reliabilities pass Martin and Bateson's (1986) acceptability level of at least 0.7.

9.3.2 Comparing average scores of the experimental and control group

An extra variable called GROUP was added when carrying out the statistical tests. This was to distinguish between students who used the CAI (i.e. experimental group) and those who were taught conventionally (i.e. control group). Students with values under this GROUP variable were classified in the following way:-

GROUP 1 = UCL control group;

GROUP 2 = Westminster control group;

GROUP 3 = Thames Valley control group;

GROUP 4 = UCL experimental group;

GROUP 5 = Westminster experimental group;

GROUP 6 = Thames Valley experimental group.

Consequently all experimental groups may be compared against all control groups or a selective comparison may be made between experimental versus control in each university.

Since two *examiners* were used mark the Marginal Costing test, each student was given two sets of scores for the pre-test and post-test (i.e. one score given by A and the other by K). To condense the two sets of scores, they were added together and divided by two (i.e. averaged). This left us with an averaged pre-test and an averaged post-test score for each student. The first aim of this research was to find out whether students had benefited from the marginal costing system as much as conventional teaching, therefore an average score was calculated for each group and an *improvement* score generated (i.e. by subtracting the pre-test from the post-test):-

	Averaged Pre-test	Averaged Post-test	Improvement Score
Experimental Group (n=23)	5.71	2.35	-3.36
Control Group (n=19)	4.87	3.79	-1.08

Table 9.4: Improvement in overall group mean scores

Clearly the experimental group improved less than the control, however, another finding is that the pre-test scores are much better than the post-test scores. The latter can be largely accounted for by a significant difference in difficulty between the two tests.

9.3.3 Independent (Unrelated) t-test to test the difference between groups

In order to test whether the marginal costing CAI was at least as effective as conventional teaching, an independent t-test was used (see appendix F). A t-test was chosen because it enables you to test if the means of the two samples differed significantly against the null hypothesis. Again a one-tailed test of significance was chosen because the alternative hypothesis was that the CAI (experimental) group was at least as effective as the conventionally taught group (i.e. control group).

Variances	t-value	df (degree of freedom)	Two-tail Significance
Equal	-2.57	40	0.014
Unequal	-2.65	39.35	0.011

Table 9.5: Independent t-test.

Looking up the t-value of 2.57 in the table of critical values of t (using 40 d.f.) we find our value is larger than the critical value of 2.423 (in the statistical table), indicating a significance level of $p < 0.01$ for a one-tailed hypothesis or $p < 0.02$ for a two-tailed hypothesis. Since the mean of the scores show that the control group had larger scores, rather more than the experimental group. In other words, the conventionally taught students demonstrated better scores than the students exposed to CAI. Consequentially, we can reject our experimental hypothesis that this CAI has an improved affect on learning compared with conventional teaching.

9.3.4 Cronbach's alpha - reliability of the attitudinal scales

(using internal consistency method)

The second aim of this research was to ascertain a number of criteria for assessing the quality of this system, using the questionnaire and especially the attitude items. Firstly, the attitude items were tested for reliability. Reliability is the extent to which a test or experiment has *consistency*, this can be either internal or external consistency. Internal consistency (or reliability) in the context of multiple-item scales is whether each scale is measuring a single idea i.e. whether it is internally consistent. Cronbach's alpha was used to measure this consistency among all the items representing attitudes to the CAI system. Cronbach's alpha essentially calculates the average of all possible split-half reliability coefficients.

The attitudinal items were grouped together under the various headings in the questionnaire (e.g. *Route Through the Program*). Then to ensure that all items were coded

in the same direction it was necessary to use the recode statement to reverse the scores on the questionnaire (all the variables that were changed were given a new variable name starting with "Z"). The result was that recoded responses that were scored 5 were changed to 1, a score of 4 became 2, score of 2 became 4 and a score of 1 became 5 (it was not necessary to recode 3, since it stays as 3). Finally they were all subjected to the SPSS test for Cronbach's alpha (the full output can be seen in appendix G). The partial output is provided in the table on the following page.

Cronbach's alpha varies between 0 and 1, and according to Wright and Fowler (1986) the test of a good reliability is if the coefficient is not less than 0.8. The test of internal reliability is important in multiple-item scales, such as the ones used in the questionnaire since it raises the question of whether each scale is measuring a single idea. Clearly from our results only three scales seem to fulfil this internal consistency. These are: (1) *the calculator facility*, (2) *comparison with other media* and (3) *accountancy knowledge/teaching*.

This is surprising especially since the items on the questionnaire were explicitly grouped together under related aspects of the system. Clearly the students seem to be judging the value of the CAI system under a number of different dimensions. Consequently, this is explored further by factor analysis. An additional observation worth making is that four of the scales showed a large degree of internal incongruity, namely *explanation (feedback) instructions*, *graphics facility*, *individual flexibility*, and *control of learning*.

Questionnaire group heading	Cronbach's alpha value	Number of items
CALCULATOR FACILITY - CALC	0.8571	3
COMPARISON WITH OTHER MEDIA - COMPAR	0.8389	11
ACCOUNTANCY KNOWLEDGE/TEACHING - ACC	0.7994	17
HELP FACILITY - HELP	0.6834	2
ROUTE THROUGH PROGRAM - ROUTE	0.6575	11
PRESENTATION INFO. ON SCREEN - PRESENT	0.6261	6
HOLISTIC SYSTEM - HOL	0.6142	10
ERROR MESSAGES AND DOCUMENTATION - ERROR	0.5714	7
RETENTION AND LEARNING - RETEN	0.5149	3
FUNCTIONS OR KEYS - FUNCTIONS	0.5030	8
COMPUTING KNOWLEDGE - COMPUT	0.4785	6
OVERALL PERCEPTION - PERCEP	0.4216	3
EXPLANATION (FEEDBACK) INSTRUCTIONS - EXPLAN	0.1808	8
GRAPHICS FACILITY - GRAPH	0.1789	4
INDIVIDUAL FLEXIBILITY - INDIVID	0.0871	5
CONTROL OF LEARNING - CONT	0.0240	4

Table 9.6: Cronbach's alpha of grouped attitude items from questionnaire.

9.3.5 Factor Analysis of attitudinal scales

Factor analysis was selected to identify any underlying concepts obtained from the questionnaire which describe key evaluation criteria. Since our results showed only three scales were internally consistent these were used in the factor analysis. Factor analysis is a statistical procedure often used to sort out survey questions into groups, called *factors* or *clusters*, for which the mean scores or percentages appear to vary together (i.e. they are

highly correlated). The usual reason for using this procedure is to condense as many of the variables in the questionnaire into a few fundamental properties, thus serving the cause of scientific parsimony. A factor analysis was carried out on the three scales (the full results can be seen in appendix H) using Kinnear and Gray (1994). The first stage was to search for common variance present in a matrix of correlations, here is the output:-

- - - - - F A C T O R A N A L Y S I S - - - - -

Analysis number 1 Listwise deletion of cases with missing values

Correlation Matrix:

	ACC	CALC	COMPAR
ACC	1.00000		
CALC	-.14949	1.00000	
COMPAR	.56287	-.24440	1.00000

Determinant of Correlation Matrix = 0.6422281

Kaiser-Meyer-Olkin Measure of Sampling Adequacy = 0.53983

Bartlett Test of Sphericity = 8.93004, Significance = 0.03024

From the correlation matrix we can see that *accountancy knowledge/teaching* has a high correlation with *comparison with other media* indicating they are measuring something in common. Moreover the determinant of the matrix is larger than 0.00001, therefore it can be assumed not to suffer from multicollinearity or singularity.

The matrix also fulfils the minimal standards for a factor analysis. The Kaiser-Meyer-Olkin measure should be greater than 0.5 for a satisfactory factor analysis to proceed and

the significance for the Bartlett test of sphericity should not exceed 0.05 (to avoid the danger that the correlation matrix is an identity matrix).

Even though some common variance was found, we were unable to carry out further factor analysis of the other scales because Bryman and Cramer (1994) reported that sample size is an important consideration of how reliable the factors emerging from this procedure. They reported that a good guideline is that there should be a minimum of five subjects per variable. Since there were twenty-six students in the experimental group, this was not enough for an additional confirmatory analysis.

9.4 APT Results

The same research design was carried out on the Application Program Tutor (APT), which is a knowledge-based tutoring system based on research done by The Electric Brain Company. Delta Performance Systems' APT is designed to develop and to teach courses for prospective users of computer applications. A marginal costing course was developed using the tutorial and exercises in the Marginal Costing Tutor and the practical part of the study was carried out at two educational institutions: De Montfort University and North Worcester Institute. All students were studying a course with a Marginal Costing as a key element.

9.4.1 Description of APT Participants

Attributes of the respondents were obtained by running a frequency command (see appendix Q). There were twenty students, of whom the majority were from De Montfort University (i.e. 12) and rest from North Worcester Institute. The former were undergraduate accountancy students on a Combined Studies course. The latter were civil servants on a short training course at North Worcester. Just over half of the students (55%) were female and exactly half were between 18 and 21 years old. Again the majority were first or second-born, which showed that they were likely to be high achievers.

When their computer experience was questioned, a large proportion (75%) reported they had a computer at home, with 70% acknowledging they could use it and 65% of them were comfortable operating it without assistance. Almost all respondents had access to a computer at work or college (95%) and clearly this was easier to use than at home since 85% said they could use it, with 80% asserting they could do it without a user manual or another person. These figures point to a better familiarity and confidence with computers, i.e. they were not naive computer users. However, there was less enthusiasm about using the computer, only 55% suggested they enjoyed using it “quite a lot”.

Much of the respondents computer education had been following school, half had never or rarely used a computer at school. Consequently, the majority (85%) had gained computer experience on training courses or through formal qualifications. Indeed a basic understanding was obtained by 25% in the form of either O level Computer Science, databases, spreadsheets, or wordprocessing. This training was not diverse since 85% admitted they couldn't even program in one language.

9.4.2 Participant's assessment of the APT

The frequency distribution of responses (see appendix Q) was examined for any pattern to the appraisal given of APT by the participants in the study. As a whole, they praised the learning system. Respondents denied emphatically that any prior computing knowledge was required to use APT. They favoured APT over books (or pen and paper) but not a lecturer's teaching. It scored better when the quality of feedback was examined, the specificity of instructions made a greater impression. The majority were confident that they would retain much of what they had learnt. Overwhelming praise was expressed for the ease with which the keys and other functions could be used. Many related a high approval rating for the presentation of screen information (a greater level than the Marginal Costing Tutor).

Crucially, some aspects favoured less well in their functionality. APT was considered to be not as flexible as the Marginal Costing Tutor. Fewer agreed on the degree of

understanding of accounting provided by this tutor. APT was found not to be helpful in terms of its error messages. There was expressive agreement that APT enabled less control of learning. Many users found the help facility wanting. The calculator facility was not viewed as positively. There was lower significance placed in the graphics capabilities of the APT.

Most respondents were uncertain whether a clear route had been provided by APT through the learning task. On the whole, APT was found to be less enjoyable to use than the Marginal Costing Tutor, but more helpful in pointing out corrections to be made. There was also ambiguity when asked about their overall perception, many were interested in using such a system but would have preferred more examples.

The responses from the open-ended questions (see the three tables in appendix R) demonstrated that the respondents strengths of the APT to be its structured teaching path and quick feedback. Its weaknesses were the lack of control and the unfriendliness of help facility. The suggested improvements to the system were better error messages and more detailed help.

9.4.3 Inferential Statistics

As in the previous study the pre-test and post-tests were marked by two different people and then compared for their inter-rater reliability (see appendix S). Again the reliabilities for both sets of tests passed the correlation acceptability threshold. The averaged pre-test and averaged post-test scores were compared for the experimental (i.e. APT) group and the control (i.e. conventionally taught) group. The computer-assisted instruction group performed much better than the traditional taught students.

	Averaged Pre-test	Averaged Post-test	Improvement Score
Experimental Group (n=10)	4.50	8.20	3.70
Control Group (n=10)	3.625	5.30	1.675

Table 9.7: Improvement in overall group mean scores.

An unrelated t-test was given to the two groups by classifying all students using a variable called GROUP under the following categories:-

GROUP 1 = North Worcester Institute experimental group;

GROUP 2 = De Montfort experimental group;

GROUP 3 = North Worcester Institute control group;

GROUP 4 = De Montfort control group.

An independent t-test, using one-tailed test of significance, was run on the control groups against the experimental groups (see appendix T). Using Greene and D'Oliveira (1993) to look up the t-value of 2.04 in the table of critical values of t (using 18 d.f.) we find our value is smaller than the critical value of 2.552 (in the statistical table), indicating a significance level of $p < 0.01$ for a one-tailed hypothesis. In other words, even though the CAI group showed better improvement than the conventionally taught students, this was not statistically significant.

9.5 Sample Size and Statistical Power

As a general rule, aiming for a large sample size (where sample is taken to be the portion of the target population upon which the research is being conducted on at one time) is taken to be a prudent strategy for the success of any research. It is certainly true that larger samples are more representative of the target population compared with smaller samples. The main reason for this is that one unusual member of a "small" sample could distort the results more than if that member was part of a "larger" sample.

Indeed, as Kerlinger (1973) reports it took Jacob Bernoulli twenty years to work out this intuitive law which states that, "Provided the members of the samples drawn independently, the larger the sample the closer the 'true' value of the population is approached", and is commonly known as the central limit theorem.

We can illustrate this with reference to a statistical measure, the "standard error" of the mean. This measure gives us some idea as to the amount of variability is present in the distribution of sample means or how likely it is that the value of any particular sample mean is in error. Large standard errors indicate great variability, whereas small ones tell us that particular sample mean is likely to be quite close to the actual population mean. The standard error of the mean is calculated by dividing the standard deviation by the square root of the sample size. The study on the APT program used a small sample size (consisting of 20 students, i.e. 10 participants in each experimental group). The standard error of the mean for the control group and the experimental group was 0.867 and 0.481 respectively. If we had increased the sample size to 60 students (and assumed the same value for standard deviation) then the standard error works out to be 0.5 and 0.278 respectively. Note that as the sample size has increased greatly, the standard error of the mean gets smaller (but only by a small amount). As Cox (1958) states the advantage of this increase in sample size is that we can draw more cogent conclusions.

In our literature review we have reported that small sample sizes are problematic in evaluation studies (see sections 3.6.2.2 and 4.4.1). Moreover acknowledging our reference in section 9.3.5, it would be advantageous to increase our sample for the purpose of an additional analysis. It is difficult to recommend an ideal minimum size (since it varies from one research design to another), however, we can report that Shute and Regian (1993) calculate that at least 30 subjects per condition are needed for simple treatment comparisons in ITS evaluation studies. They emphasise that acquiring not only the right "number" but also the right "type" of subjects is paramount. Whilst a more general recommendation of 110 or more participants was referred to by Altman (1991). However this issue is not that simple, as we can see from the calculation of standard error of mean (which is inversely related to the square root of the sample size). Also increasing the

sample size has a diminishing effect on improving the estimate of the population mean (Martin et al. 1986).

Kerlinger (1973) offers a word of caution on the subject of estimates of error. He states that we cannot tell how much this standard error is biased. Kerlinger also relates an example from a researcher who in studying large numbers of Air Force pilots found that estimates of standard errors were sometimes considerably off. Indeed, he asserts "No one can give hard and fast rules....Calculate statistics and act as though they were 'true', but always maintain a certain reserve toward them,". His best advice in this regard is to use random sampling and keep observations independent, if at all possible. Random sampling is simply a way of ensuring, as much as possible, that each member of a population has an equal chance of being selected for the sample. Thus guaranteeing in the long run the sample taken represents the population. While we do not have the opportunity to do this topic justice here, it certainly worth examining references such as Kidder et al. (1981), which provide a good discourse on choosing appropriate sampling strategies.

Standard error of the mean has its origins in sampling error (arising because a finite sample rather than the whole population is measured) but it is only one possible type of error (or uncontrolled variance). An alternative source of error that may arise in a study is called measurement error and this is the combined error that results from inevitable imperfections and variability in the process of measurement. All of these possible "error variances" have to be reduced in order for the researcher to detect the effect on dependent variable(s) after manipulating the independent variable(s). We will discuss the importance of statistical power, but for now error variability plays a significant part in research effects as Hays (1981) states, "It follows that controlled experiments in which there is little 'natural' variation in the materials observed can attain statistical power with relatively few observations, while those involving extremely variable material may require many observations to attain the same degree of power". As we have discussed previously, this increase in excessive control can result in the reduced external validity of the finding, so a balance has to be achieved. Lastly, it is worth noting that many of the assumptions of the most common measures of error rely on the normal distribution (see Hays, 1981 for further information).

We also need to consider the sensitivity or power of a significance test. This relates to making hypothesis testing. Conventionally researchers need to decide between mutually exclusive options, i.e. accepting the null hypothesis (i.e. there are no noticeable differences between the conditions) or accepting the alternative (experimental) hypothesis. However, this judgement can be fraught with danger (see Kidder et al. 1981) since four outcomes are possible. The first two are making the correct decision i.e. either there are no differences between conditions due to the "treatment" or there are differences between conditions due to the effectiveness of the treatment. However the remaining two outcomes are errors in decision making known as Type I error and Type II error. Briefly, Type I error occurs when we reject the null hypothesis more readily than we should, simply by accepting a less rigorous standard (i.e. a lower level of statistical significance) and is a false positive. Type II error is the opposite in that we accept the null hypothesis even in situations where the differences between the conditions is sufficiently large to signify a research effect (i.e. false negative). Alpha is denoted as the probability of making a Type I error and beta is probability of making a Type II error. Power is normally defined as $1 - \beta$ and is the probability of correctly rejecting the null hypothesis when the alternative is true. Standardised (or z) scores can give an indication of the statistical power. Snodgrass (1977) provides a good example of how increasing sample size results in increase in z scores (and consequently statistical power).

Cox (1958) describes that in order to make the correct decision (therefore avoid Type I or Type II errors) we must examine the "sensitivity" or power of a significance test. Significance levels provide an indication of whether this difference is large enough to have occurred by chance alone. Power is a good concept to aim for and so it is worth maximizing it as much as you can. There are a number of factors that affect power of a statistical test, including factors outside the investigator's control (i.e. under nature's control). One example of the latter is where the effect of "treatment" is not very different from the "non-treatment" condition, e.g. in a problem solving performance study, where the test is affected by feedback. Other items more readily under the investigator's control are: choice of an appropriate statistical test, significance level and preference for a one-tailed test (i.e. predicting the direction of difference in advance). Adelman (1992) adds five main ways of increasing statistical power: (1) increase the sample size, (2) increase

the reliability of the study (e.g. increasing the reliability of the measurement instruments), (3) give careful consideration to the research design, (4) give careful consideration to the research question and (4) consider the importance of the "effect size". Finally, Cox (1958) puts power in context when he says "Power is important in choosing between alternative methods of analysing data and in deciding on an appropriate size of experiment. It is quite irrelevant in the actual analysis of data". Indeed, the fact that the results are "statistically" significant does not necessarily mean that it is "socially" significant.

9.6 Conventional Teaching of the Control Groups

The intelligent computer assisted instruction packages that were used in this study were outlined in the procedure of this study. The main ITS used was the Marginal Costing Tutor which was designed to teach a topic in accountancy. Its key feature was its highly interactive and problem-solving teaching strategy. Following a number of on-screen tutorials the user tests their understanding by solving a number of randomly generated question scenarios. Instant feedback is provided by advice tailored to that part of the solution (for more information on the subject domain and the tutoring characteristics please refer to sections 8.2 and 8.3 of this thesis). The other ITS used in this research work was the aptly named Application Program Tutor. The APT is a tutor shell (i.e. no subject domain is included) and therefore requires subject knowledge to be specifically developed for it to "teach". This was one of the main reasons for it not being used initially (i.e. not only would research participants have to be sought but a course, comparable to a conventionally taught one, would have to be designed and piloted before any serious work could be considered). There are also implications for the usefulness of the findings on the interface design (please refer to section 8.2 for a fuller account of the APT).

The principal idea for the research was to determine whether the Marginal Costing Tutor could do better against an incumbent instruction method using the target user group and how the Tutor could be improved. The practical was carried out at the following educational institutions for the Marginal Costing Tutor: University College London, Thames Valley University and University of Westminster. The latter of these studies involved postgraduates. However, the former studies all involved undergraduate students,

the University College participants were working towards an additional option (outside their principal course) in management accountancy. All three courses had a component of lectures and course work. University College London students were predominately taught by one lecturer with the "chalk and talk" style as well as hand outs. Students from Thames Valley University had several teachers and invariably relied on overhead projection slides as well as directed reading. Whilst the University of Westminster postgraduates had a mixture of tutorials and lectures.

Looking at the APT studies, the De Montfort University students were all attending an undergraduate combined studies in accounting course. Of course they were familiar with a Marginal Costing ITS (though not in its older version that we had used, instead they were more aware of the marginal cost part of the Byzantium ITT set of modules). They were taught by a combination of methods including lectures and tutorials. In contrast the remaining APT study consisted of mainly mature students actively seeking a diploma on a financial management course in a Higher Education college.

CHAPTER TEN

CONCLUSIONS

"An education isn't how much you have committed to memory, or even how much you know. It's being able to differentiate between what you do know and what you don't. It's knowing where to go to find out what you need to know; and it's knowing how to use the information once you get it." [William Feather, 1889 - 1981] (15;20;8;5;2;1;20;20)

This chapter examines the results obtained in this study and compares them with findings from other similar investigations. In addition there is an analysis of the experience and conclusions reached from this work and the implications for the advancement of this research area.

10.1 Objectives of this Research

The literature review established there were several key reasons for being very troubled by a major obstacle in the field of computer-aided instruction (CAI). This considerable gap in the literature can be summed up by the following phrase - a high development rate but little enduring effectiveness.

It is characterised by the following:-

- the stress on developing intelligent CAI as research vehicles at the expense of integrating them as educational tools into the mainstream of schools, higher education and training institutions
- the shortage of commitment and patience to even develop workable versions of CAI systems (consequently most are destined to remain prototypes)

- a disenchantment by educators and users with overoptimistic claims of the benefits of CAI
- negligence to invest heavily in specialist expertise (i.e. multidisciplinary development teams)
- the scarcity in specialised software tools for construction and in the necessary skills for maintenance
- a requirement to utilise a high degree of innovative computing resources (as opposed to standard or portable technologies)

This has resulted in a culture where construction is still a one-off enterprise and current trends demonstrate that there will continue to be:-

- an overabundance of poorly designed educational software
- a deficiency in shared evaluation theories, tools and practices
- a reliance one-off specific (i.e. case-study type) evaluations
- an inadequate documentation of common standards and successful operational systems

until radical change is implemented in this research area.

Throughout this work, we described the importance of evaluation for the further advancement of this field, both to improve the overall standard of CAI systems produced and the wider acceptance of this technology outside the narrow confines of research laboratories. But, the contemporary absence of a long-term paradigm and suitable evaluations techniques (to facilitate this process) are delaying this advancement. At this point, we introduced an alternative approach, to the normal models associated with the field of evaluation, which was selected from the philosophy of quality. However, before

we reiterate the benefits of this new direction, we need underline the lessons learnt about the nature and purpose of good evaluations:-

- Evaluation serves as an important link between *past* and *future* activities (Rogers et al., 1992),
- Evaluation requires the establishment of the *goals of educational software* (Underwood and Underwood, 1990),
- Evaluation needs to be focused on the *instructional effectiveness* of software as well the technical characteristics (Reiser and Kegelmann, 1994),
- Evaluation must be viewed as an invitation to *learn about people and systems* and involves systematically gathering the *worth* and *value* of an instructional enterprise (Winne, 1993),

Evaluation is, at best a time-dependent activity and at worst is an exercise in extravagant data amassing. It does, however, follow critical incidents of the development cycle, i.e. obtaining snapshots of a system's health both early and late in development. Our reason for embracing quality management in our new approach (over and above current evaluation methods) is the strong emphasis on making requirements explicit and using this activity to support a process of continuous improvement. Another important distinction between evaluation and quality is that while the former involves product assessment at a specific instances, the latter's drive for improvement extends well beyond the product to include the process and continues to be active throughout the development cycle (consequently the process as well as the product become key factors in the improvement strategy). The two approaches are best viewed as complementary rather than equivalent, so that while evaluation is merely an audit or review conversely, quality is a set of high-level strategies completely influencing a system's or organisation's practices.

Not only did we draw on quality to inform the philosophy underlying this thesis but we also utilised it in a more pragmatic way, most notably in the way the study focused

attention on the end-users (inclusion of the intended users of the system and not the so-called experts) in the system's revision and construction of a questionnaire which sought "fitness for purpose" information for the system. Indeed for the questionnaire, we used the Software Quality Metrics model to allow the customer to assess the final product. This model is appropriate to deriving external attributes of a product.

This approach is not that extraordinary, since Davis and Venkatesh (1996) used a similar attitudinal scales, as this study, to evaluate information technology. Their paper describes the Technology Acceptance Model (TAM) which was developed by IBM Canada, Ltd in the mid-1980s to evaluate the market potential for a variety of PC-based applications. Since then TAM has become one of the most powerful models explaining IT acceptance and use. It has been very successfully applied to the study of a wide range of issues related to acceptance. They report that "While TAM does not cover all possible factors affecting product assessment, empirical research has revealed that the constructs of perceived usefulness and perceived ease of use compare favorably with alternative approaches such as: realism of expectations...,motivational force...,user satisfaction...,user information satisfaction and user involvement...,value...,theory of planned behavior...,and theory of trying".

From what we know, there are two basic ways to improve on quality once a measurement is made of a system currently performs, either it can be checked against its own best performance as part of an ongoing process or it can be compared against a common standard. Using this approach, we were able to specify the operational definitions of *value* and *effectiveness* for our study. These were expressed as the two aims of this research namely which key *characteristics* of the system are significant in the assessment of a quality teaching system and whether a CAI system was at least as *effective* as conventional teaching. Indeed this was corroborated from our examination of the literature where Winne (1993) states that two forms of worth and value are relevant in evaluating an educational system: (1) to promote students' achievements and (2) the qualities of an engineered computing system.

Consequently, the purpose of this work was to investigate a number of measures to assess the *value* and *effectiveness* of *operational* computer-aided instruction systems with intended users.

This means that our major objectives were to:-

- utilise an already working system,
- find out if it is effective,
(i.e. is it at least as good as the equivalent conventional method?)
- elicit key properties that define it as valuable system
(i.e. discovering lessons for future systems)
- develop a procedure which could accomplish these goals easily and effectively, for evaluators.

10.2 The research findings

10.2.1 Marginal Cost Tutor findings

We obtained an operational system (i.e. a Marginal Costing Tutor) and carried out a study on its intended users (i.e. university accountancy students) using a well crafted questionnaire and a series of competency tests.

- We found that the results demonstrated that conventional teaching had a *greater* improvement on learning than students using CAI.
- In addition, three consistent attitude dimensions emerged as key valuation judgements used by end users, that is:-
 - (1) *accountancy knowledge/teaching*,
 - (2) *comparison with other media*, and
 - (3) *calculator facility*.

The characteristics of the participants showed that:-

- All of them were studying accountancy with a component of marginal costing at university. More than three-quarters were undergraduates and the remainder postgraduates. The majority were between 18 and 22 years old and approximately half were male students. A large number of them were first or second-born indicating they were high achievers.
- Most, if not all, students had access to computers and a working knowledge how to use them, with over half professing that they enjoyed using them quite a lot. Very few had used computers on a regular basis at school. The majority, however, had received some form of computer training since school (most notably in spreadsheets and wordprocessors) but only a handful admitted to being ardent programmers. There were no major differences between the comparison groups in computer training.

Exposure to the Marginal Costing Tutor resulted in the following response:-

- On the whole, the reaction to the accounting system was favourable. Users liked this system because it was fun to use and could be used interactively. Overall they confirmed the designer's assertion that it did not require a lot of computing knowledge. In this regard a well-designed screen design and a careful arrangement of lessons was seen as a successful factor, consequently the respondents attributed the system as "quick and easy to use", "self explanatory" and having a "clear and logical" program design.
- Another important aspect was the number of examples and exercises which encouraged them to become rapidly involved and gain a basic grounding in marginal costing. One clue of how this system was being used came from comparison with other media where they rated it better than a textbook method of learning because it was interactive and resembled one-to-one tutoring. These key features resulted in a suitable pace and direction of learning, comfortable control of keys and facilities (especially of the calculator and the graphs) and a good grounding in the use of marginal costing information. *As a result users*

were happy to be given this system either to revise or for assignments since it increased their specific ability to solve marginal costing problems.

- In spite of the above response, users recognised it could be improved a lot more. A major advancement would be to make the instructions more elaborate and reduce the ambiguity of error responses. Most users were satisfied with the feedback once a response had been entered. Moreover, a small number indicated a preference for a suggested route through the program. A complementary suggestion was to incorporate the exercises (problems) into the tutorials instead of having two separate sections and to include more examples. A less popular proposal was to make the screen design more attractive (through better graphics and colours) although we believe this to have been influenced by their experience of a Graphical User Interface (i.e. Windows) environment.

10.2.2 Replication with Application Program Tutor

Another study was conducted with a different system (i.e. Application Program Tutor or APT) using the same questionnaire and same tests from the first study.

- We found that the students using the computer assisted instruction (i.e. APT) produced significantly larger improvements than those taught conventionally. However, this result was not statistically significant.

A profile of the participants showed that:-

- 60% of students were from De Montfort on a Combined Studies Accountancy course and the remainder were civil servants at North Worcester Institute on Financial Management diploma course. 50% were between 18 and 21 years old (comparable to the previous study). Approximately half were female students and over half (65%) were 1st or 2nd born (again, comparable to the previous study).

- Three-quarters confessed to having a computer at home, and almost all of these could use it, most of them were comfortably without assistance. There is a similar, if not better, picture at work or college where 95% had a computer.

- Half had never or rarely used a computer at school, yet 85% had received some form of computer training or education (a quarter of whom within O Level computer science, database, spreadsheet, or wordprocessing) but many admitted they couldn't even program in one language. So the computer experience was quite specific, most of it package based.

Assessment of the APT was mixed:-

- Participants were generally positive of APT in that: prior computing knowledge was not required, they favoured it over books (or pen and paper), quality of feedback was rated highly, as was the ease with which the keys and other functions could be used, they were impressed by the specificity of instructions, they would retain much of what they had learnt, the presentation of screen information was endorsed and the structured teaching path was considered a notable strength.

- However, APT was considered less flexible than the Marginal Costing Tutor and this was symptomatic of their view about the rest of the system: APT was found to be less helpful in terms of its error messages, it enabled less control of learning, many found the help facility, graphics capabilities and the calculator facility of little use. These statements were reflected in the weaknesses of the system.

- On the whole, APT was found to be less enjoyable to use than the Marginal Costing Tutor, but more helpful in pointing out specific corrections to be made. The participants seemed to be rating the problems as originating from the interface which they asked more improvements be made, notably better error messages and more detailed help.

10.2.3 Previous Marginal Cost Tutor studies

Wilkinson-Riddle and Patel (1993) had developed a prototype system called "Human Emulation Software" to teach the subject of Marginal Costing within an area of Management Accounting Education. The prototype software was able to provide a clear mathematical relationship of marginal costing through an interactive approach of problem setting and problem solving. There have been several studies conducted using this Marginal Costing Tutor with accountancy students (a few of them in the case-study style of evaluation).

- Tonge and Johnston et al. (1993) examined the experiences of using the Marginal Costing Tutor at two different institutions (i.e. University of Teesside and De Montfort University) and three different student groups (i.e. first year undergraduate students on modular business course, a course for women returning to higher education, and staff and students on a combined studies accounting course).

- The first study at Teesside involved lengthy feedback from the lecturers and student responses using a questionnaire after more than 3 weeks of usage. The results showed that lecturers saw interactivity as a major strength and students were positive about the ease of use and the method of learning. The second study at De Montfort involved questionnaires and interviews following 6 weeks of course interaction (again competence was not tested). The students on three different years found the program "reasonable" to use for practical problem solving especially for revision purposes. But they wished more lecturer access, particularly at the beginning of use. They concluded that the interaction proved to be the real strength of the system (rather than the "books on screens" available as management accounting packages) and this development should be encouraged to evolve.

- Dungworth and Russell (1994) focused on the exposure of Marginal Costing Tutor on three distinct groups: (i) traditional A level entry students, (ii) mature returnees to education, and (iii) overseas students. Data was collected three methods: (a) questionnaire for students, (b) interviews of staff and students, and (c) observations by staff and

students. The findings for traditional and overseas students showed that after initial apprehension and period of adjustment (5 weeks) they were able to deal with most complex problems. However, mature students preferred teacher centred learning. Two conclusions were reached from the work, firstly students need training in using CAL and they require support facility to answer any questions on the package and topic.

- Tonge, Beacham and Gallagher (1994) reviewed the progress in application of information technology on accounting education over the previous two decades, including an evaluation of a “Human emulating software” which uses a management accounting application (i.e. Marginal Costing Tutor). The paper reported on detailed trials which took place at the University of Teesside on first year accounting students. Sixty one students were exposed to three weeks of intensive exercises including paper based examples and the software’s random generator. The outcome of the evaluation was measured by questionnaire, performance at the end of the module (especially noting their choice of questions), and a number of interviews (on a small group of students).

- Respondents were positive about the usefulness of many features of the software, in particular they liked the worked examples and the practice examples, considered the software easy to use, and liked the individual instruction and feedback. All students gained in knowledge of the subject (those with prior knowledge showing the greatest improvement) and more students answered the “break even” question than any other. The conclusion from the evaluation was that the software is an effective learning vehicle, but its potential has not been fully reached due to a lack of real embedding in the module course. The curriculum aspects of CBL are paramount for its successful application for undergraduate users.

10.2.4 Management Accounting Module

The success of this Marginal Costing Tutor and the research interest in the system led to the setting up of a six university consortium led by De Montfort University to bid for Teaching and Learning Technology Programme (TLTP) funding from the Higher Education Funding Councils of the UK. This led to the development of a new set of

Intelligent Tutoring Tools (each of which is mixed-initiative tutoring system). Each tool has been designed to cover one elementary topic of the curriculum at the "application level" but full knowledge of the subject at the "advance application level". This fresh approach to computer integration in learning environments is known as the Byzantium Model. Three modules for the teaching of introductory accounting were developed in 1993-95 under the TLTP Byzantium Model at De Montfort University, i.e. Capital Investment Appraisal; Absorbtion Costing and Marginal Costing.

The results of our research work on the Marginal Costing Tutor were fed back into the De Montfort team. So that while the new TLTP model was being developed and refined a number of significant features, that were improved, resulted from direct feedback of the results of our study. For example, our study highlights that one of the key strengths of the Marginal Costing Tutor was the large amount of examples and exercises. As a consequence of this, this capability of generating random problem scenarios was enhanced in the new system to include both the creation of totally new questions but also favourite questions used by a student. Another strength was its self-explanatory style ("told me answer was wrong and how to work it out"), however some students in our study had also seen this as a weakness. Therefore more work was carried out on this feature.

A "Dynamic Messaging System" component was created in the inference engine which generated a context sensitive message whenever the student enters value. If the student enters a correct answer, they are informed the value is right and the student is allowed to proceed further. For values that are incorrect but can be worked out from data already present there is a four level help facility: (i) hint of incorrect value, (ii) suggestion of best possible relationship, (iii) suggestion of the data for calculation and (iv) correct answer. For those where enough information has not been provided by the student to calculate the variable, there is suggested best possible relationship to calculate corresponding intermediate variable.

This was an advance on the Marginal Costing Tutor where the student was told if they made a mistake they were simply provided with enough information to calculate the correct value. Therefore, although the some students made progress through the Marginal

Costing Tutor their learning was not context-driven. These students subsequently complained about this aspect as the weakness of the system i.e. "More elaborate and self-explanatory instructions, not enough information to carry on, not user friendly" or "Poor feedback (more praise)". In our study the "pop-up calculator" was selected as a favourite feature of the interface. This was critically improved in the new system to match the interface and to be interactive with other controls of the application, for example enabling values to be uploaded to and from the editbox onto the screen with a single click of the mouse. Please refer to Kinshuk, 1996 for an extensive description of the new system and its evolution.

Subsequently, evaluative research was carried out using the questionnaire developed by our research but with this updated version of the software (upgraded to the Windows version, and very different from the MS-DOS version we had used in our study) now part of a set of Management Accounting Modules. It is also worth noting that the new software was subjected to the same comparative research style that was demonstrated in our study.

- Kinshuk (1996) focused on the effectiveness of the Teaching and Learning Training Programme (TLTP) Byzantium CAL modules (i.e. Capital Investment Appraisal, Absorption Costing and Marginal Costing). Exactly the same interface design was used in all three modules used in the evaluation. A number of variables were varied/measured: (a) effectiveness of different CAL modules, (b) subject topic characteristics, (c) different curriculum structures of universities, (d) effects of learning style, (e) background attributes of the students. The evaluation techniques used in the study (i) subject based evaluation (structured questionnaire and observations by author and respective tutors) and (ii) experimental evaluation. The study was conducted across 4 new universities and 2 as traditional universities.

- The structured questionnaire provided the strengths and weakness of the system. A majority of students reported the strengths as: user interface of the package, error handling and ease of navigation. They found the packages easy to use and understand. This was primarily due to the layout of the screens. *More importantly the feedback from the*

student's responses on our questionnaire supplied recommendations to improve features of the package: calculator facility, interface layout, and saving routines of the package.

- Observational records demonstrated that students did not have any problems in using the packages and could find their way through the programs without any external help. Students found it easy to learn the subject matter with the help of the packages. No significant differences between traditional teaching and CAL teaching methods. Learning styles had no effect.
- Students agreed that the package does not required prior knowledge of accounting or computing. Students were less agreeable that the software compares well with traditional forms of learning or whether the software provides helpful error messages and it is easy to follow. No significant associations were found with user opinions and learning styles. Students without prior computer training were less supportive of the statement that no prior knowledge of accounting was required. However, no significant difference in performance was found between students with or without confidence in operating computers or those who enjoyed (or didn't enjoy) computers. Moreover, the performance of students was not associated with any level of previous computer training. *The results using our questionnaire were published by Patel and Kinshuk (1996).*

The following is a comparative summary between our study and Kinshuk's work:-

- Systems evaluated:

- (i) Marginal Costing Tutor (MS-DOS version and written in Clipper language)
- (ii) Capital Investment Appraisal, Absorbtion Costing and Marginal Costing Modules (Windows version and written in Visual C++)

- Educational Institutions used in evaluative study:

- (i) 3 UK institutions (Thames Valley University, University of Westminster and University College London)
- (ii) 6 UK institutions (4 were ex-polytechnics and 2 were traditional universities)

- Number of Participants:

- (i) 42 students

- (ii) phase I 40 conventionally taught

- 40 taught using CAL

- phase II 513 students

- Focus of work:

- (i) Development of evaluation measures and comparative study with traditional teaching

- (ii) Evaluation of "new" Accountancy CAL and the effectiveness of CAL in relation to student's learning styles.

10.3 Conclusions drawn

- Firstly, there is no current model of evaluation for ITS (apart from the broad attempt that Bento, 1990 suggested). This is key point. The absence of the theoretical underpinnings means that researchers must draw heavily on other existing and closely-related fields to draw in relevant abstract issues (such as eliciting quality metrics).

When researchers understand that end-users are extracting characteristics (such as consistency of the ITS in *comparison with other media* i.e. how close or different this ITS is to the incumbent form of instruction) of the system that make the ITSs useful or successful (or even select ones that make the system unusable or unsuccessful) then practical advantages can be gained by comparing the original design or objective of the system and making changes in the interface, the features (like the calculator facility) or even the speed of the inference engine (just like the enhancements to Marginal Costing Tutor).

Juristo et al. (1997) have produced a framework but only after literally years of groundwork and working in a field which is only relatively recently begun to work together to address such issues. We have certainly a long way to go in the educational

technology area before such a model can be developed further. It will have to be developed by a bottom-up approach since the ground work still needs to be done.

- Looking back at our research, the questionnaire did prove as fruitful as we had predicted. Indeed from our carefully crafted questionnaire we found a lot of useful information about the system's strengths and weaknesses, especially how to improve it (which was one of the underlying objectives of this work, see section 1.2.2). In addition, the Likert Scale was found to be a successful way of obtaining student assessments. Three different systems have been evaluated using this questionnaire (each system with a totally different interface) the Marginal Costing Tutor, Application Program Tutor and the TLTP Byzantium CAL modules. The last of these systems is a commercial product which is undergoing successful development (in part due to our work). Therefore this questionnaire (in conjunction with the pre-test and post-test) was useful in providing answers to what standards of worth and value are relevant in evaluating that system (see sections 4.4.4 for a fuller information on the questions i.e. educational and engineering).

- As we have already indicated the study did show that similar findings were obtained on the key features of older Marginal Costing Tutor compared to the new version of the Marginal Costing Module. This suggests that the users did perceive common (though not the same) beneficial and disagreeable features of both systems. Consequently, since both systems are from the same "family of systems" we can conclude the questionnaire does seem to elicit the external attributes of the system design. However as far as the APT, our evaluation study also managed to pick up the inflexibility (i.e. lack of interactivity) of the APT compared with the Marginal Costing Tutor and the reduction in range of topics (reflecting the limitations of the APT screen capture). Moreover the limitations on teaching style on APT e.g. no helpful advice (compared with an "adaptive" Marginal Costing Tutor) did little justice to enable a fair comparison with conventional teaching. Although this is not a final confirmation of the usefulness and external validity of the questionnaire. It does prove that the questionnaire can be used with other system evaluations and is open to be improved and extended for other systems (e.g. a revised version was used by Kinshuk for his three separate modules).

- We have to say, however, that two main obstacles posed serious problems in this research. Firstly, there was a large problem in the unavailability of educational software which could be considered to be *fully operational* as well as being used in education.
- Secondly, the task of carrying out an evaluation within a real applied teaching environment resulted in a large amount of the time was spent searching for suitable end users and arranging facilities for the study. Consequently, this affected the size of the study sample as well as the hurried choice of a marginal costing test (since an earlier test was found to be solvable without using marginal costing procedures). This is not a common problem since developers of software can easily obtain access to the target users, however, independent evaluators are compelled to rely on the generosity of educators for access to their students. In this regard the APT study was found severely wanting. While it is acknowledged that larger sample sizes are desirable in any future research in this area, this sole issue has not detracted from the value of this type research (indeed there are some researchers who value just using the single-person case study). What instead we need to focus upon how potent is the finding in the predictive sense (as well as its usefulness in moving the field on).

Hays (1981) puts it best when he reports "Highly analyzed and significant results are often confused with good results. But overemphasizing the role of statistical analysis and statistical significance in research is like confusing the paintbrush with the painting. This form of statistical inference is a valuable tool in research, but it is never the arbiter of good research". We must not forget that we have used a systematic approach to research. Whilst there is no such thing as an ideal design and there are numerous pitfalls (outlined in section 2.6.1) we are confident that we have obtained the most useful information that we can (given the classic tradeoff, see sections 4.4.1).

- One of the major findings was that conventional teaching produced better test scores than the Marginal Costing system. This was not so surprising given the relatively short time the students had to use the CAI. Another possible explanation for these results can clearly be seen by examining the low post-test scores for both groups, although this test was meant to be slightly harder it may have been too difficult. This software has been

successfully used by a number of educational institutions (including De Montfort University, Liverpool Polytechnic and Middlesex University) for more than four years in place of tutorials and evidence from the past exam results and current evaluative work (Kinshuk, 1996) show it is an effective educational system. Therefore, it is reasonable to argue that our results were influenced by the sensitivity of test we used and the exposure period to the CAI. This was substantiated when greater exposure time to the APT in our study produced better overall scores in both tests (so the tests seem to be sensitive to the period of learning, unsurprisingly).

- More substantive work has been carried out by Kulik (1994) with larger sample sizes. Using the evaluation literature to identify representative results using meta-analysis of individual studies, he found that overall computer tutoring compares favourably well with results of other innovations. His results showed that "the average effect of computer-based instruction was to raise examination scores by .35 standard deviations, or from the 50th to the 64th percentile". However, he did benefit from examining groups of studies with multivariate techniques where the effect size can be more apparent than comparisons of individual studies.

There was overwhelming evidence that the Marginal Costing Tutor was successful with students:-

- They used it like an interactive book and were happy with its ease of use. Two items dominated improvements, a better on-line help mechanism and a more sophisticated arrangement of lessons (a sizeable number requested a refresher facility half-way through the exercises).
- It is also true that a number of responses were not so straightforward. For instance, students favoured the feedback because they immediately knew whether they were right or wrong but they were uncertain whether the explanations showed them clearly *why* they had got the wrong answer (further work is needed here to obtain the reasons for this, e.g. we may require a more precise set of questions to make the reasons for this clear).

- The results for the APT suggested that the expert shell nature of the system coloured the response from the users (the APT uses other applications to derive its screen information i.e. does not have a domain-specific knowledge). Many were familiar with computers and used them both at home at work or college, as a consequence they did not find it enjoyable to use (due lack of interactivity) but did improve using the teaching material provided. In fact, the APT students improve substantially more than the conventional taught students, although it is worth noting this was not statistically significant. We are pleased with the result of the evaluation since it provides proof that the questionnaire is able to derive the key individual strengths and weaknesses in a CAI system, in despite of the interface (in fact it takes into account the superficial significance of the interface). In summary, the APT study has confirmed the suitability of this evaluative approach. John Cook (1991) has been involved in the evaluation of APT for many years.

Our research contribution has been to:-

- establish a substantial grounding in the field of educational technology (including historical perspective, the salient issues and the current problems and obstacles) for example a CAI approach demands far greater investment of resources before any output is obtained

- extend evaluation by applying quality management philosophy to the design and evaluation practice, for instance

(a) advocating an open approach to development, testing, standards and maintenance (see section 2.6.1 and section 6.5.1);

(b) apply a well-known quality doctrine - design and test for end-user i.e. customer-oriented (see Basili, 1995 in section 6.5.2 and section 8.3.1 and Winne, 1993 in section 4.4.4);

(c) learning exercise to improve - drew on lessons from piloting questionnaire (see section 8.3.3).

- draw heavily on existing expertise in the evaluation literature to report what to do and what not to do, using models of evaluation and examples of evaluation studies, for example

(a) tried to steer an even course between the dangers of internal and external validity (see Duncan, 1993 and Ransdell, 1993 in section 2.6.1; Shute and Regian, 1993 in section 4.4.1);

(b) optimising for the best models and styles of evaluation (see Hopkin, 1989 in section 5.4 and Knapper, 1980 in section 5.6).

The specific achievements of this thesis has been to:-

- show there is a huge gap literature both in the theory of CAL evaluation and in its customary practice. The only evaluation model for ITSs found during the research period as Bento (1990) and unfortunately it proved to be inappropriate to evaluate the Marginal Costing Tutor. At the time of writing this area attracts too little funding or focus to generate improvement.

- show that it is possible to conduct an evaluation which not only determines whether an instructional system meets its targets but also gathers information to gauge its worth or value (i.e. part determination of requirements and part setting of standards). The evaluation style we have chosen (see section 5.6) enables other researchers to carry this out even when resources are tight and there is an urgent need for key data to suggest improvements. Remember any chosen strategy must: (i) be easy to use otherwise it will be ignored due to money or time constraints, (ii) yield results that be used effectively and (iii) be capable

of being revised or adapted e.g. a questionnaire. Since our work meets these criterion they are present a powerful argument against avoiding evaluation.

- provide a new way of tackling long-term issues in this research area by using the model of quality assurance in real-life educational setting (Iqbal et al., 1998 "Use of Quality Metrics in User-Oriented Evaluation Standards Development for Computer Assisted Instruction Systems"). This research document is available in the conference proceedings or details are currently available on-line at the following Internet address:

<http://ctiwebct.york.ac.uk/LTSNCiPAbstracts/cip98.htm>

- Although we have used the SQM model (see section 6.5.2 described by Basili, 1995) for the above research, we do "not" have a preference to any specific philosophy of quality management. However we have stressed in the whole of chapter six that quality management as a whole should be applied to the software in this artificial intelligence field and the reasons are clearly given in that and preceding chapters. This philosophy has had a direct effect (not only by encouraging the reusability or lengthening longevity of a real-life system) through the feedback of this work into the design of a new version of the Marginal Costing Tutor, as well as the creation of a questionnaire which has been adapted to test a new version of the Tutor (to develop if further) and has been used to obtain responses from the APT which is a totally different type of system. This is the real philosophy of quality i.e. the encouragement of a developer culture willing to learn and a long-term commitment to improving its craft, tools and techniques. In addition, the development of this questionnaire has elicited three consistent quality metrics (which are now in the public domain) and can be used and tested for use in future work by other evaluators.

- The second point worth understanding is that the general approach taken by this work "initiates" (i.e. not complete) this new path by considering one aspect, i.e. the elicitation of "some" quality metrics from a real-life working ITS and from its intended user group. It is important to note that this is a beginning and not a completion. Furthermore if we were selecting one model to stress for current and future work we could do no better than

promote the Software Quality Metrics model by Basili (1995). This has been described in Iqbal et al. (1998) as well as sections 6.5.2, 10.1 and supported by Juristo (1997).

Other achievements of this thesis has been to:-

- devise a successful questionnaire which is being used by a company (E.W. Fact and Educational Software Products) improve the development of its commercial educational computer assisted instruction system
- obtain two operational systems which we could use in a field study (despite large number of problems - Professor Tim O' Shea said there were NO operational ITSs)
- obtain clear feedback on the comparative usefulness of the system with conventional teaching (with a novel and ethical design strategy, i.e. variation on the before-after two-group design, never reported in the literature)
- obtained results showing how to improve a commercial system (Marginal Costing Tutor) with an extensive report to other developers (thus fulfilling the statement in section 2.6.1 that the benefits to be accrued by reporting more of our systems cannot be stated)
- carried out an experimental style study in a real teaching environment (despite not having access to the originally intended users)
- developed an instrument (i.e. a well-crafted questionnaire) which was used by other evaluators (at De Montfort University and now is in the public domain i.e. to be useful to others)
- published a number of papers with the developers of the Marginal Costing Tutor to promote the impetus behind this effort to other potential evaluators, for all current and future references see the on-line publication link at the Internet address given below:-

<http://fims-www.massey.ac.nz/~is/staff/kinshuk.html>

- developed and published a classification system for useful evaluation techniques (Iqbal, Oppermann et al., 1999 "A Classification of Evaluation Methods for Intelligent Tutoring Systems"). The basis of our current investigation on the suitability of evaluation techniques to the ITS field was first described in section 3.7 of this thesis. It has been designed to help other evaluative researchers to select appropriate techniques for their study. It is elaborated further in the above research paper to extend the evaluation field wider and bring it to the attention of a larger audience (including reference to the work in this thesis), it is also currently available on-line on the Internet at the following address: http://fims-www.massey.ac.nz/%7Ekinshuk/papers/se99_eval_methods.pdf
- suggested a brand new method of tackling long-term issues (not just issues specific to one CAL system) using the framework of quality assurance (only just being started in the expert system field, see Juristo, Mate and Tovar 1997).

We must emphasize that the work mentioned in the last statement above is the start of a new chapter in teaching system evaluation but there are a few signs, that like the Knowledge Based Systems area of Juristo et al. (1997), researchers are beginning this type of work.

10.4 Future Lessons Learnt

What improvements do we recommend for the future of CAI and its evaluation?

- avoid the temptation to adopt the develop-test-throwaway culture evident in much of AI research, instead encourage a shift of culture to successive development which encourages software reuse (e.g. object-oriented design, suggested by Seffah, 1997)
- urge all designers of educational systems to set targets of what their software ought to accomplish after being used by a chosen user (including period of exposure and improvement in educational achievement)

- encourage the move towards evaluation as a key development activity and away from the attitude that evaluation is taboo or difficult (by making it easier to use and test a range of relevant evaluation methods, tools and techniques)
- discourage the tendency to conduct single case studies type evaluations (which take longer to conduct than the rest of the development and accumulate large amount of idiosyncratic data) and instead exploit evaluation methods and techniques from related disciplines which harvest practical worth or value for future development
- be guided by the five goals addressed by Winne (1993) when conducting a high calibre evaluations
- persuade more independent evaluations since designers cannot always recognise improvements to be made to their own software and promote more public awareness when systems are operationally successful (i.e. which serve as useful exemplars)

The most pressing line for future research is the need to:-

- develop evaluation methodologies which cope with the newest of technologies, where knowledge is not constrained to the linear-document form. For instance multimedia or Internet where knowledge is packaged in non-linear ways and allows the user to explore it by self-determined linkages.

There is a need that some of the lessons that we have learned in this research about the effective use of evaluation to be applied to all aspects of computer-aided instruction. This successive style of development and evaluation seems to be the only way to increase the effectiveness and usefulness of such systems, resulting in better software, wider usage and a longer shelf-life.

APPENDIX A

Questionnaire

SURVEY OF A TUTORING SYSTEM

BY

ARIF M. IQBAL

PLEASE GO THROUGH THE QUESTIONS IN THIS DOCUMENT
AND TRY TO ANSWER THEM AS CAREFULLY AS POSSIBLE.

MOST OF THE QUESTIONS RELATE TO THE
COMPUTER PROGRAM YOU HAVE JUST USED.

ALL INFORMATION WILL BE TREATED WITH THE STRICTEST CONFIDENCE

PAGE 1

A. BIOGRAPHICAL DATA

On the following few sections are a list of questions about yourself and your familiarity with computers.

Please respond either by **ticking** the appropriate box on the right-hand side of the page or by **writing** the relevant answer on the dotted lines provided.

1. Please indicate your name here.....

2. Your University.....

3. Your department or course.....

4. What is your age (in years)?.....

5. What is your present occupation?

Undergraduate student

☐ 5.

Postgraduate student

☐

If other, please specify.....

6. Please specify your gender.

Male

☐ 6.

Female

☐

7. In which order were you born in your family?

First-born

☐ 7.

Second-born

☐

Third-born

☐

If other, please specify.....

B. CURRENT EXPERIENCE WITH COMPUTING

8. (a) Do you have a computer at HOME?

Yes ☐ 8 a.

No ☐

(b) If YES, do you use it?

Yes ☐ 8 b.

No ☐

(c) If YES, on the whole, can you operate it without help of a user manual or other person?

Yes ☐ 8 c.

No ☐

9. (a) Do you have access to a computer at COLLEGE/WORK?

Yes ☐ 9 a.

No ☐

(b) If YES, do you use it?

Yes ☐ 9 b.

No ☐

(c) If YES, on the whole, can you operate it without help of a user manual or other person?

Yes ☐ 9 c.

No ☐

10. How much do you enjoy using a computer?

Not applicable ☐ 10.

Very much ☐

Quite a lot ☐

Unsure ☐

Not much ☐

Not at all ☐

C. PREVIOUS EXPERIENCE WITH COMPUTING

11. How often did you use a computer at school?

Daily ☐ 11.
 Weekly ☐
 Monthly ☐
 Rarely ☐
 Never ☐

12. (a) Have you received any computer related education/training?

Yes ☐ 12 a.
 No ☐

(b) Which type(s) of education/training did you receive?

Not applicable ☐ 12 b.

Education -

'O' level Computer Science

☐ 'O'

'A' level Computer Science

☐ 'A'

HND Computer Science

☐ HND

BTEC Computer Science

☐ BTEC

BSc Computer Science

☐ BSc

Software Packages -

Database

☐ Data

Spreadsheet

☐ Spread

Word processing

☐ Word

If other, please specify.....

13. (a) Can you program in any computer languages?

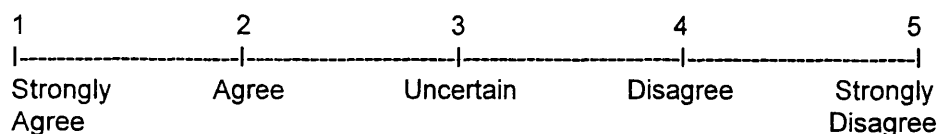
Yes ☐ 13 a.

No ☐

(b) How many different computer languages can you program in?

Not applicable ☐ 13 b.

Number of languages.....



D. ASSESSMENT OF THE TUTORING SYSTEM

Given below are a number of statements about the marginal costing system, please indicate your response for each statement, **using the scale above**. For example, if you feel that you **strongly agree** about the statement seven, then write **1** in the box next to statement seven. However if you **strongly disagree** about statement number seven then write **5** in the box. There are no right or wrong answers, please give the response closest to your own feelings.

COMPUTING KNOWLEDGE

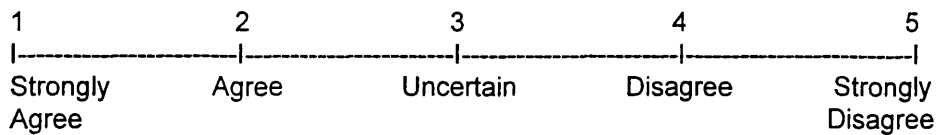
COMPUT

14. This program does not require any prior knowledge of computing.		14.
15. I didn't learn anything because I was spending all my time getting to know how to use a computer.		15.
16. After using this system I became more confident about learning from a computer.		16.
17. It was difficult enough to tackle the accountancy problems without having to fight with the computer.		17.
18. The system can be used sufficiently well without any manuals.		18.
19. Normally I find this type of subject very simple, but I couldn't see what the computer wanted me to do.		19.

COMPARISON WITH OTHER MEDIA

COMPAR

20. This system is a better teacher than most lecturers.		20.
21. Having a lecturer close by, so that I could ask any questions, would have been a big improvement.		21.
22. I liked understanding marginal costing with this system, because it's like being on a one-to-one basis with your tutor.		22.
23. I was left with a lot of unanswered questions after using the system.		23.
24. This form of teaching was really clear, because unlike a lecturer teaching, the program does not miss out a lot of steps.		24.



25. It is easier to get involved solving accounting problems using this system than in a classroom tutorial. 25.

26. I prefer an accountancy book because you can flip it backwards and forwards. 26.

27. It's much easier to understand than reading an accountancy book because it's interactive. 27.

28. It was terrific to be able to discover the relationships between variables, which I never got from books. 28.

29. I would rather learn accountancy by working out problems using pencil and paper. 29.

30. I like doing accountancy this way because you can see how it's done, rather than having to ask somebody to explain it. 30.

INDIVIDUAL FLEXIBILITY

INDIVID

31. This system is only appropriate for advanced students of accounting. 31.

32. The computer got harder just as I got better at the problems. 32.

33. I understood what I had to do, but the system persisted in following its long and laborious steps. 33.

34. The system acts like a human tutor because it adjusts its advice according to the individual's needs and progress. 34.

35. I found it helpful to jump into the more demanding material without having to go through the easier stuff. 35.

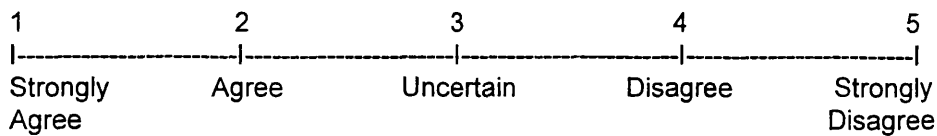
ACCOUNTANCY KNOWLEDGE/TEACHING

ACC

36. I didn't need any previous knowledge of accounting to use the system. 36.

37. I have gained a good introduction to the concepts of marginal costing. 37.

38. I think the program is only useful for people who are familiar with accountancy and want to improve. 38.



39. At the beginning I was not sure what marginal costing meant, but once I got the hang of it, I became quite an expert. 39.
40. It is difficult to learn the basics of marginal costing with this system. 40.
41. The system gives a first class grounding in the use of marginal costing information. 41.
42. Most people would find it hard to work out marginal costing problems using this system. 42.
43. The system allows you to fully test your understanding of marginal costing. 43.
44. At the end I still did not know anything about marginal costing. 44.
45. My knowledge of accounting has been greatly increased by using such a system. 45.
46. After using this system I can easily use my knowledge to work out marginal costing for most information given about a firm. 46.
47. I would find it easy to use this system to teach someone else about marginal costing. 47.
48. The computer has helped me by guiding me through the learning. 48.
49. I would like to learn other accounting topics using this system. 49.
50. The computer system allowed me to learn an accurate picture of marginal costing and correct my mistakes. 50.
51. It allowed me to think about the relationships behind marginal costing. 51.
52. Even now, I do not understand the relationships between the different parts of marginal costing. 52.

ERROR MESSAGES AND DOCUMENTATION

53. The computer messages were good enough to help me when I got stuck. 53.
54. The error messages were unhelpful. 54.

ERROR

1	2	3	4	5
----- ----- ----- -----				
Strongly	Agree	Uncertain	Disagree	Strongly
Agree				Disagree

55. I liked the program because it allows me to make mistakes, so that it can advise me what to do. 55.

56. The program should not allow you to enter any incorrect values. 56.

57. The program is so unforgiving, it doesn't allow you to do anything else, till you provide the correct answer. 57.

58. There should be an error message if I enter an incorrect negative or positive numbers. 58.

59. I like the way it told me precisely what I did wrong. 59.

EXPLANATION (FEEDBACK) INSTRUCTIONS

EXPLAN

60. The system only provides good feedback once you have entered an answer. 60.

61. The system didn't give enough praise for the correct answer. 61.

62. The system helped me with every step I took in solving the problem. 62.

63. Through the feedback I immediately knew whether I had got it right or wrong. 63.

64. The explanations given by the program are hard to understand. 64.

65. The explanations showed clearly why I had got the wrong answer. 65.

66. The program gives individual feedback to each student depending upon how far they have progressed to the solution. 66.

67. A clearer explanation of the aim of the exercise and the information needed to complete it would have been much better. 67.

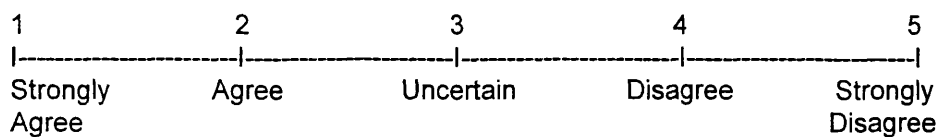
ROUTE THROUGH PROGRAM

ROUTE

68. I never knew what all the items on the menu meant. 68.

69. The instructions given on how to proceed through a tutorial were unclear. 69.

70. I like learning like this, because you can check over things and not make the same mistakes. 70.



71. Sometimes I found it hard to keep track which bits I had done. 71.
72. It allowed me to solve a problem in any order. 72.
73. The system is unable to recognise a different route to the solution. 73.
74. I would have found it more helpful to be given a suggested route through the program. 74.
75. The prompts, asking me to enter text, weren't clear from the screen. 75.
76. I want to be able to stop the exercise and go back to the tutorial, and then return where I left off. 76.
77. On-screen tutorials are difficult to follow. 77.
78. It provided little understanding of what you are doing or how to do it. 78.

CONTROL OF LEARNING

CONT

79. I felt the system allowed me to work at my own pace and direction. 79.
80. I found the system was too slow because I could do the exercises much more quickly with pencil and paper. 80.
81. It was too quick, I would like more control of how much time I was given to solve the problems. 81.
82. I learnt a lot about marginal costing by trying different things with the system. 82.

RETENTION AND LEARNING

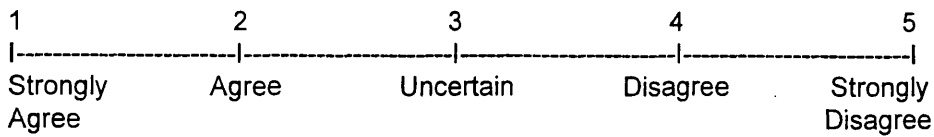
RETEN

83. I am sure I will retain most of what I have learnt. 83.
84. I learnt most from this system, by trying to answer the various problems. 84.
85. If I use this program again, I could solve similar problems much quicker. 85.

HELP FACILITY

HELP

86. I would prefer more help from the system. 86.



87. The help facility should have given a few hints, rather than the correct answer. ☐ 87.

FUNCTIONS OR KEYS

FUNCTIONS

88. I found it very difficult to know which keys corresponded to the command I wanted. ☐ 88.

89. When proceeding through the program I found I was clear which options I was allowed to select. ☐ 89.

90. I hated some of the option keys, I sometimes got lost because of them. ☐ 90.

91. When I was going through the program, pressing the same key would produce a different command. ☐ 91.

92. It was easy to exit from a particular route through the program. ☐ 92.

93. I couldn't skip some parts of the exercise. ☐ 93.

94. It was simple to delete any answer which I knew was wrong. ☐ 94.

95. I wasn't allowed to restart with the same data. ☐ 95.

CALCULATOR FACILITY

CALC

96. The calculator facility was easy to use. ☐ 96.

97. I found it hard to use the calculator to transfer the answer to the program. ☐ 97.

98. Using the calculator is frustrating - it doesn't do what it's supposed to do. ☐ 98.

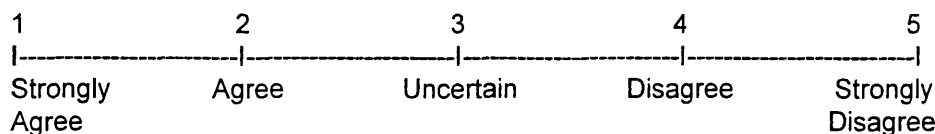
GRAPHICS FACILITY

GRAPH

99. I found the graph very useful. ☐ 99.

100. It was difficult to read the information from the graph. ☐ 100.

101. The graphical facility allows you to compare easily the effect of different decisions on costs. ☐ 101.



102. I would have enjoyed it better if there had been more graphics, colours and sounds. ☐ 102.

PRESENTATION INFO. ON SCREEN

PRESENT

103. The screen layout meant that it was easy to communicate with the computer. ☐ 103.

104. Sometimes I was confused by the large amount of information on the screen. ☐ 104.

105. The way the problem is presented on the screen makes it easy to work out the solution. ☐ 105.

106. There was not enough information on the screen for me to solve the problems. ☐ 106.

107. The screen was attractive and clear. ☐ 107.

108. The colours on the screen were very confusing. ☐ 108.

HOLISTIC SYSTEM

HOL

109. I really liked using this system because it was fun to use. ☐ 109.

110. I enjoyed it but it could be improved a lot more, so I would be tempted to use it again. ☐ 110.

111. I have used a computer program like this before, but this one is a lot better. ☐ 111.

112. I simply used the computer to read what is on the screen. ☐ 112.

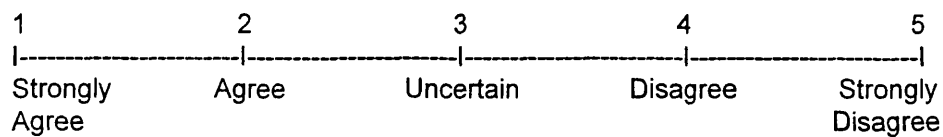
113. The program went on too long. ☐ 113.

114. The more you use it, the easier it becomes. ☐ 114.

115. I would be happy, if I could be given assignments using this program. ☐ 115.

116. The program did not manage to convince me that I could learn something by using it. ☐ 116.

117. I liked the way it helped me when I was in trouble. ☐ 117.



118. It is a good way to revise something I have forgotten. 118.

OVERALL PERCEPTION

PERCEP

119. Most students are not interested in using such a system. 119.

120. I solved the problems but I was unclear why the answer was correct. 120.

121. I would like to have more examples. 121.

PLEASE TURN OVER

122. Apart from any of the issues discussed previously,
what other features did you find contributed to the:-

(a) STRENGTH of this system

.....

.....

(b) WEAKNESS of this system

.....

.....

123. If there was **ONE** outstanding improvement you could recommend
to this system, what would it be?

.....

.....

PLEASE GO BACK AND CHECK YOU HAVE ANSWERED ALL THE QUESTIONS,

AND

PLEASE MAKE SURE YOU HAVEN'T LEFT ANY RELEVANT BOX OR LINE BLANK.

THANK YOU VERY MUCH FOR YOUR CO-OPERATION.

APPENDIX B

Marginal Costing Tests

MANAGEMENT ACCOUNTING RESEARCH

BY

ARIF M. IQBAL

MARGINAL COSTING NUMBER 1

**YOU ARE REQUIRED TO ANSWER ALL PARTS
OF THE QUESTION IN THIS PAPER.**

TIME ALLOWED: 10 MINUTES.

ALL INFORMATION WILL BE TREATED WITH THE
STRICTEST CONFIDENCE BY THE RESEARCHER.

N.B. THESE RESULTS WILL HAVE NO EFFECT
WHATSOEVER ON YOUR COURSE WORK.

BEFORE ATTEMPTING THIS PAPER PLEASE
READ THE QUESTION CAREFULLY.

CALCULATORS MAY BE USED BUT PLEASE
SHOW ALL YOUR WORKING CLEARLY.

PAGE 1

Q1 . A company makes a single product with a sales price of £10 and a marginal cost of £6.
Fixed costs are £60,000 p.a.

Calculate:

- a) Number of units to break even,
- b) Sales at break-even point,
- c) Contribution/Sales ratio, and
- d) What number of units will need to be sold to achieve a profit of £20,000 p.a.?

WRITE YOUR ANSWERS BELOW AND ON THE NEXT PAGE

Please write your name and course details below and then answer the question using the space given. Show **CLEARLY** how you arrived at the answer.

NAME OF STUDENT.....

COURSE OF STUDENT.....

A1 .

MANAGEMENT ACCOUNTING RESEARCH

BY

ARIF M. IQBAL

MARGINAL COSTING NUMBER 2

**YOU ARE REQUIRED TO ANSWER ALL PARTS
OF THE QUESTION IN THIS PAPER.**

TIME ALLOWED: 10 MINUTES.

**ALL INFORMATION WILL BE TREATED WITH THE
STRICTEST CONFIDENCE BY THE RESEARCHER.**

**N.B. THESE RESULTS WILL HAVE NO EFFECT
WHATSOEVER ON YOUR COURSE WORK.**

**BEFORE ATTEMPTING THIS PAPER PLEASE
READ THE QUESTION CAREFULLY.**

**CALCULATORS MAY BE USED BUT PLEASE
SHOW ALL YOUR WORKING CLEARLY.**

PAGE 1

Q2 . A company makes a single product with a sales price of £10 and a marginal cost of £6.
Fixed costs are £60,000 p.a.

The company is considering a proposal to launch their product in a new market segment. The finance director has drawn up an action plan, which will be necessary if the launch is to be successful.

This comprises of:-

1. A reduction of 10% to the current price.
2. A special advertising campaign at a cost of £10,000.
3. Purchase of a new high quality component at an extra cost per unit of £1.50.

If this plan is implemented the finance director forecasts sales of 60,000 units.

Based on the above scenario calculate the following:

- a) Number of units to break even,
- b) Sales at break-even point,
- c) Contribution/Sales ratio,
- d) What number of units will need to be sold to achieve a profit of £30,000 p.a.?
- e) The margin of safety.

WRITE YOUR ANSWERS BELOW AND ON THE NEXT PAGE

Please write your name and course details below and then answer the question using the space given. Show CLEARLY how you arrived at the answer.

NAME OF STUDENT.....

COURSE OF STUDENT.....

A2 .

MANAGEMENT ACCOUNTING RESEARCH

BY

ARIF M. IQBAL

MARGINAL COSTING NUMBER 3

**YOU ARE REQUIRED TO ANSWER ALL PARTS
OF THE QUESTION IN THIS PAPER.**

TIME ALLOWED: 10 MINUTES.

**ALL INFORMATION WILL BE TREATED WITH THE
STRICTEST CONFIDENCE BY THE RESEARCHER.**

**N.B. THESE RESULTS WILL HAVE NO EFFECT
WHATSOEVER ON YOUR COURSE WORK.**

**BEFORE ATTEMPTING THIS PAPER PLEASE
READ THE QUESTION CAREFULLY.**

**CALCULATORS MAY BE USED BUT PLEASE
SHOW ALL YOUR WORKING CLEARLY.**

PAGE 1

Q3 . A company makes a single product with a sales price of £10 and a marginal cost of £6.
Fixed costs are £60,000 p.a.

The company is considering a proposal to launch their product in a new market segment. The finance director has drawn up an action plan, which will be necessary if the launch is to be successful.

This comprises of:-

1. A reduction of 20% to the current price.
2. A special advertising campaign at a cost of £20,000.
3. Purchase of a new high quality component at an extra cost per unit of £0.75.

If this plan is implemented the finance director forecasts sales of 120,000 units.

Based on the above scenario calculate the following:

- a) Number of units to break even,
- b) Sales at break-even point,
- c) Contribution/Sales ratio,
- d) What number of units will need to be sold to achieve a profit of £30,000 p.a.?
- e) The margin of safety.

WRITE YOUR ANSWERS BELOW AND ON THE NEXT PAGE

Please write your name and course details below and then answer the question using the space given. Show **CLEARLY** how you arrived at the answer.

NAME OF STUDENT.....

COURSE OF STUDENT.....

A3 .

APPENDIX C

Frequency Distribution of Questionnaire Responses

UNIV university of subject

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
thames valley univer	tv	12	28.6	28.6	28.6
university college l	ucl	21	50.0	50.0	78.6
university of westmi	uw	9	21.4	21.4	100.0
		-----	-----	-----	
Total		42	100.0	100.0	

Valid cases 42 Missing cases 0

AGE age of subject

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	18	4	9.5	9.5	9.5
	19	10	23.8	23.8	33.3
	20	4	9.5	9.5	42.9
	21	9	21.4	21.4	64.3
	22	4	9.5	9.5	73.8
	23	2	4.8	4.8	78.6
	25	3	7.1	7.1	85.7
	27	2	4.8	4.8	90.5
	28	1	2.4	2.4	92.9
	30	1	2.4	2.4	95.2
	32	1	2.4	2.4	97.6
	35	1	2.4	2.4	100.0
		-----	-----	-----	
Total		42	100.0	100.0	

Valid cases 42 Missing cases 0

OCCUP occupation of subject

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
UNDERGRADUATE	1	33	78.6	78.6	78.6
POSTGRADUATE	2	9	21.4	21.4	100.0
		-----	-----	-----	
Total		42	100.0	100.0	

Valid cases 42 Missing cases 0

SEX gender of subject

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
male	1	23	54.8	54.8	54.8
female	2	19	45.2	45.2	100.0
		-----	-----	-----	
Total		42	100.0	100.0	

Valid cases 42 Missing cases 0

BIRTH_OR In which order were you born in your family?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
1st-born	1	14	33.3	33.3	33.3
2nd-born	2	12	28.6	28.6	61.9
3rd-born	3	9	21.4	21.4	83.3
4th-born	4	4	9.5	9.5	92.9
5th-born	5	2	4.8	4.8	97.6
6th-born	6	1	2.4	2.4	100.0
		-----	-----	-----	
Total		42	100.0	100.0	

Valid cases 42 Missing cases 0

HOME_COM Do you have a computer at HOME?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	25	59.5	59.5	59.5
no	2	17	40.5	40.5	100.0
		-----	-----	-----	
Total		42	100.0	100.0	

Valid cases 42 Missing cases 0

HOME_USE Do you use the home computer?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	20	47.6	47.6	47.6
no	2	5	11.9	11.9	59.5
not applicable	3	17	40.5	40.5	100.0
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 42 Missing cases 0

HOME_OPE On the whole, can you operate it without help of a user manual or other person?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	17	40.5	43.6	43.6
no	2	5	11.9	12.8	56.4
not applicable	3	17	40.5	43.6	100.0
	-1	3	7.1	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 39 Missing cases 3

COLL_COM Do you have access to a computer at COLLEGE/WORK?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	41	97.6	97.6	97.6
no	2	1	2.4	2.4	100.0
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 42 Missing cases 0

COLL_USE Do you use the college computer?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	37	88.1	88.1	88.1
no	2	4	9.5	9.5	97.6
not applicable	3	1	2.4	2.4	100.0
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 42 Missing cases 0

COLL_OPE On the whole, can you operate it without help of a user manual or other person?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	29	69.0	72.5	72.5
no	2	10	23.8	25.0	97.5
not applicable	3	1	2.4	2.5	100.0
	-1	2	4.8	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 40 Missing cases 2

ENJOY_CO How much do you enjoy using a computer?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
not applicable	1	2	4.8	4.8	4.8
not much	3	5	11.9	11.9	16.7
unsure	4	10	23.8	23.8	40.5
quite a lot	5	22	52.4	52.4	92.9
very much	6	3	7.1	7.1	100.0
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 42 Missing cases 0

COMP_SCH How often did you use a computer at school?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
never	1	8	19.0	19.0	19.0
rarely	2	14	33.3	33.3	52.4
monthly	3	5	11.9	11.9	64.3
weekly	4	14	33.3	33.3	97.6
daily	5	1	2.4	2.4	100.0
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 42 Missing cases 0

COMP_EDU Have you received any computer related education/training?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	29	69.0	69.0	69.0
no	2	13	31.0	31.0	100.0
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 42 Missing cases 0

TYPE_EDU Which type(s) of education/training did you receive?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
other	0	3	7.1	7.1	7.1
not applicable	1	9	21.4	21.4	28.6
Databases	7	1	2.4	2.4	31.0
Word processing	9	2	4.8	4.8	35.7
N/A, Spread	18	1	2.4	2.4	38.1
N/A, Word	19	1	2.4	2.4	40.5
BSc, Other	60	1	2.4	2.4	42.9
Spread, Word	89	5	11.9	11.9	54.8
Data, Spread, Word	789	7	16.7	16.7	71.4
O Level, Data, Spread,	2789	8	19.0	19.0	90.5
BSc, Data, Spread, Wor	6789	1	2.4	2.4	92.9
Data, Spread, Word, O	7890	3	7.1	7.1	100.0
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 42 Missing cases 0

COMP_LAN Can you program in any computer languages?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	11	26.2	26.2	26.2
no	2	31	73.8	73.8	100.0
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 42 Missing cases 0

NO_COMP How many different computer languages can you program in?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
not applicable	0	31	73.8	75.6	75.6
one language	1	5	11.9	12.2	87.8
two languages	2	2	4.8	4.9	92.7
three languages	3	2	4.8	4.9	97.6
four languages	4	1	2.4	2.4	100.0
	-1	1	2.4	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 41 Missing cases 1

PRIOR_KN This program does not require any prior knowledge of computing.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	14	33.3	60.9	69.6
uncertain	3	5	11.9	21.7	91.3
disagree	4	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

NOTLEARN I didn't learn anything because I was spending all my time getting to know how to use a computer.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
disagree	4	12	28.6	52.2	56.5
strongly disagree	5	10	23.8	43.5	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

CONFIDEN After using this system I became more confident about learning from a computer.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	11	26.2	47.8	52.2
uncertain	3	7	16.7	30.4	82.6
disagree	4	3	7.1	13.0	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

FIGHT It was difficult enough to tackle the accountancy problems without having to fight with the computer.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	4.8	8.7	8.7
uncertain	3	1	2.4	4.3	13.0
disagree	4	13	31.0	56.5	69.6
strongly disagree	5	7	16.7	30.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

OUT_MANU The system can be used sufficiently well without any manuals.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	14	33.3	60.9	69.6
uncertain	3	5	11.9	21.7	91.3
disagree	4	1	2.4	4.3	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

WHAT_TDO Normally I find this type of subject very simple, but I couldn't see what the computer wanted me to do.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	2	4.8	8.7	13.0
uncertain	3	1	2.4	4.3	17.4
disagree	4	13	31.0	56.5	73.9
strongly disagree	5	6	14.3	26.1	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

BET_TEAC This system is a better teacher than most lecturers.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	7.1	13.0	13.0
agree	2	4	9.5	17.4	30.4
uncertain	3	9	21.4	39.1	69.6
disagree	4	5	11.9	21.7	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

LEC_CLOS Having a lecturer close by, so that I could ask any questions, would have been a big improvement.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	5	11.9	21.7	21.7
agree	2	8	19.0	34.8	56.5
uncertain	3	6	14.3	26.1	82.6
disagree	4	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ONE_ONE I liked understanding marginal costing with this system, because it's like being on a one-to-one basis with your tutor.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	13	31.0	56.5	60.9
uncertain	3	6	14.3	26.1	87.0
disagree	4	2	4.8	8.7	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

UNANSW_Q I was left with a lot of unanswered questions after using the system.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	3	7.1	13.0	13.0
uncertain	3	6	14.3	26.1	39.1
disagree	4	10	23.8	43.5	82.6
strongly disagree	5	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ALL_STEP This form of teaching was really clear, because unlike a lecturer teaching, the program does not miss out a lot of steps.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	7.1	13.0	13.0
agree	2	5	11.9	21.7	34.8
uncertain	3	11	26.2	47.8	82.6
disagree	4	3	7.1	13.0	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

CLASSTUT It is easier to get involved solving accounting problems using this system than in a classroom tutorial.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	5	11.9	21.7	21.7
agree	2	9	21.4	39.1	60.9
uncertain	3	3	7.1	13.0	73.9
disagree	4	5	11.9	21.7	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

BOOK I prefer an accountancy book because you can flip it backwards and forwards.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	3	7.1	13.0	13.0
uncertain	3	5	11.9	21.7	34.8
disagree	4	13	31.0	56.5	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

I_ACTIVE It's much easier to understand than reading an accountancy book because it's interactive.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	8	19.0	34.8	34.8
agree	2	8	19.0	34.8	69.6
uncertain	3	6	14.3	26.1	95.7
disagree	4	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

REL_VAR It was terrific to be able to discover the relationships between variables, which I never got from books.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	9.5	17.4	17.4
agree	2	11	26.2	47.8	65.2
uncertain	3	4	9.5	17.4	82.6
disagree	4	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

PEN_PAP I would rather learn accountancy by working out problems using pencil and paper.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	4.8	8.7	8.7
uncertain	3	9	21.4	39.1	47.8
disagree	4	9	21.4	39.1	87.0
strongly disagree	5	3	7.1	13.0	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

SEE_HOW I like doing accountancy this way because you can see how it's done, rather than having to ask somebody to explain it.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	13	31.0	56.5	65.2
uncertain	3	3	7.1	13.0	78.3
disagree	4	4	9.5	17.4	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ADV_ACC This system is only appropriate for advanced students of accounting.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	2.4	4.3	4.3
uncertain	3	4	9.5	17.4	21.7
disagree	4	14	33.3	60.9	82.6
strongly disagree	5	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

COM_HARD The computer got harder just as I got better at the problems.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	7	16.7	30.4	30.4
uncertain	3	7	16.7	30.4	60.9
disagree	4	9	21.4	39.1	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

LONGSTEP I understood what I had to do, but the system persisted in following its long and laborious steps.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	2.4	4.3	4.3
uncertain	3	8	19.0	34.8	39.1
disagree	4	14	33.3	60.9	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

IND_NEDS The system acts like a human tutor because it adjusts its advice according to the individual's needs and progress.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	9	21.4	39.1	43.5
uncertain	3	5	11.9	21.7	65.2
disagree	4	6	14.3	26.1	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

JUM_DEMD I found it helpful to jump into the more demanding material without having to go through the easier stuff.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	13	31.0	56.5	65.2
uncertain	3	7	16.7	30.4	95.7
disagree	4	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

PREV_ACC I didn't need any previous knowledge of accounting to use the system.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	9.5	17.4	17.4
agree	2	8	19.0	34.8	52.2
uncertain	3	4	9.5	17.4	69.6
disagree	4	6	14.3	26.1	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

GOOD_INT I have gained a good introduction to the concepts of marginal costing.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	7.1	13.0	13.0
agree	2	14	33.3	60.9	73.9
uncertain	3	5	11.9	21.7	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ONLY_ACC I think the program is only useful for people who are familiar with accountancy and want to improve.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	3	7.1	13.0	21.7
uncertain	3	6	14.3	26.1	47.8
disagree	4	10	23.8	43.5	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

BECM_EXP At the beginning I was not sure what marginal costing meant, but once I got the hang of it, I became quite an expert.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	4	9.5	17.4	17.4
uncertain	3	13	31.0	56.5	73.9
disagree	4	4	9.5	17.4	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

DIFF_BAS It is difficult to learn the basics of marginal costing with this system.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	2	4.8	8.7	8.7
disagree	4	20	47.6	87.0	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

GROUNDIN The system gives a first class grounding in the use of marginal costing information.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	14	33.3	60.9	69.6
uncertain	3	7	16.7	30.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

WORK_OUT Most people would find it hard to work out marginal costing problems using this system.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	4.8	8.7	8.7
uncertain	3	6	14.3	26.1	34.8
disagree	4	12	28.6	52.2	87.0
strongly disagree	5	3	7.1	13.0	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

FULLTEST The system allows you to fully test your understanding of marginal costing.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	17	40.5	73.9	73.9
uncertain	3	5	11.9	21.7	95.7
disagree	4	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

NO_THING At the end I still did not know anything about marginal costing.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	4.8	8.7	8.7
uncertain	3	1	2.4	4.3	13.0
disagree	4	10	23.8	43.5	56.5
strongly disagree	5	10	23.8	43.5	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

KN_INCRE My knowledge of accounting has been greatly increased by using such a system.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	11	26.2	47.8	52.2
uncertain	3	5	11.9	21.7	73.9
disagree	4	5	11.9	21.7	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

EASY_KNO After using this system I can easily use my knowledge to work out marginal costing for most information given about a firm.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	10	23.8	43.5	47.8
uncertain	3	9	21.4	39.1	87.0
disagree	4	1	2.4	4.3	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

TEACELSE I would find it easy to use this system to teach someone else about marginal costing.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	11	26.2	47.8	56.5
uncertain	3	8	19.0	34.8	91.3
disagree	4	1	2.4	4.3	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

GUID_ME The computer has helped me by guiding me through the learning.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	19	45.2	82.6	87.0
uncertain	3	1	2.4	4.3	91.3
disagree	4	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

OTHER_TO I would like to learn other accounting topics using this system.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	8	19.0	34.8	34.8
agree	2	8	19.0	34.8	69.6
uncertain	3	5	11.9	21.7	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ACCU_PIC The computer system allowed me to learn an accurate picture of marginal costing and correct my mistakes.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	9.5	17.4	17.4
agree	2	17	40.5	73.9	91.3
uncertain	3	1	2.4	4.3	95.7
disagree	4	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

RELA_BEH It allowed me to think about the relationships behind marginal costing.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	13	31.0	56.5	65.2
uncertain	3	6	14.3	26.1	91.3
disagree	4	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

NOT_RELA Even now, I do not understand the relationships between the different parts of marginal costing.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
uncertain	3	6	14.3	26.1	30.4
disagree	4	10	23.8	43.5	73.9
strongly disagree	5	6	14.3	26.1	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

HELPSTUC The computer messages were good enough to help me when I got stuck.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	9.5	17.4	17.4
agree	2	16	38.1	69.6	87.0
uncertain	3	2	4.8	8.7	95.7
disagree	4	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

MESS_HEL The error messages were unhelpful.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	2.4	4.3	4.3
uncertain	3	4	9.5	17.4	21.7
disagree	4	12	28.6	52.2	73.9
strongly disagree	5	6	14.3	26.1	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

MAK_MIST I liked the program because it allows me to make mistakes, so that it can advise me what to do.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	9	21.4	39.1	39.1
agree	2	12	28.6	52.2	91.3
uncertain	3	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

NOT_CORR The program should not allow you to enter any incorrect values.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	2	4.8	8.7	17.4
uncertain	3	3	7.1	13.0	30.4
disagree	4	8	19.0	34.8	65.2
strongly disagree	5	8	19.0	34.8	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

PROG_UNF The program is so unforgiving, it doesn't allow you to do anything else, till you provide the correct answer.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	2	4.8	8.7	17.4
uncertain	3	5	11.9	21.7	39.1
disagree	4	10	23.8	43.5	82.6
strongly disagree	5	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ERR_MSG There should be an error message if I enter an incorrect negative or positive numbers.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	9.5	17.4	17.4
agree	2	11	26.2	47.8	65.2
uncertain	3	5	11.9	21.7	87.0
disagree	4	2	4.8	8.7	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

PREC_WRG I like the way it told me precisely what I did wrong.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	6	14.3	26.1	26.1
agree	2	12	28.6	52.2	78.3
uncertain	3	3	7.1	13.0	91.3
disagree	4	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

GOOD_FED The system only provides good feedback once you have entered an answer.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	11	26.2	47.8	56.5
uncertain	3	8	19.0	34.8	91.3
disagree	4	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ENGH_PRA The system didn't give enough praise for the correct answer.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	5	11.9	21.7	26.1
uncertain	3	6	14.3	26.1	52.2
disagree	4	7	16.7	30.4	82.6
strongly disagree	5	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

HELP_STP The system helped me with every step I took in solving the problem.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	12	28.6	52.2	56.5
uncertain	3	6	14.3	26.1	82.6
disagree	4	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

IMM_FEED Through the feedback I immediately knew whether I had got it right or wrong.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	7	16.7	30.4	30.4
agree	2	14	33.3	60.9	91.3
uncertain	3	1	2.4	4.3	95.7
disagree	4	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

EXPL_UND The explanations given by the program are hard to understand.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	2.4	4.3	4.3
uncertain	3	5	11.9	21.7	26.1
disagree	4	16	38.1	69.6	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

CLE_EXPL The explanations showed clearly why I had got the wrong answer.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	7	16.7	30.4	39.1
uncertain	3	9	21.4	39.1	78.3
disagree	4	5	11.9	21.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

IND_FEED The program gives individual feedback to each student depending upon how far they have progressed to the solution.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	7	16.7	30.4	34.8
uncertain	3	12	28.6	52.2	87.0
disagree	4	3	7.1	13.0	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

AIM_EXER A clearer explanation of the aim of the exercise and the information needed to complete it would have been much better.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	5	11.9	21.7	30.4
uncertain	3	10	23.8	43.5	73.9
disagree	4	6	14.3	26.1	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ALLMEANT I never knew what all the items on the menu meant.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	4	9.5	17.4	21.7
uncertain	3	4	9.5	17.4	39.1
disagree	4	12	28.6	52.2	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

INST_CLR The instructions given on how to proceed through a tutorial were unclear.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	2	4.8	8.7	13.0
uncertain	3	6	14.3	26.1	39.1
disagree	4	12	28.6	52.2	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

SAME_MIS I like learning like this, because you can check over things and not make the same mistakes.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	7.1	13.0	13.0
agree	2	12	28.6	52.2	65.2
uncertain	3	7	16.7	30.4	95.7
disagree	4	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

HARD_TRC Sometimes I found it hard to keep track which bits I had done.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	3	7.1	13.0	17.4
uncertain	3	4	9.5	17.4	34.8
disagree	4	14	33.3	60.9	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ANY_ORD It allowed me to solve a problem in any order.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	6	14.3	26.1	26.1
agree	2	10	23.8	43.5	69.6
uncertain	3	3	7.1	13.0	82.6
disagree	4	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

DIFF_ROU The system is unable to recognise a different route to the solution.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	3	7.1	13.0	21.7
uncertain	3	12	28.6	52.2	73.9
disagree	4	6	14.3	26.1	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

SUGG_ROU I would have found it more helpful to be given a suggested route through the program.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	9	21.4	39.1	47.8
uncertain	3	7	16.7	30.4	78.3
disagree	4	5	11.9	21.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

PROM_CLR The prompts, asking me to enter text, weren't clear from the screen.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	4.8	8.7	8.7
uncertain	3	5	11.9	21.7	30.4
disagree	4	14	33.3	60.9	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

RET_LEFT I want to be able to stop the exercise and go back to the tutorial, and then return where I left off.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	7	16.7	30.4	30.4
agree	2	11	26.2	47.8	78.3
uncertain	3	4	9.5	17.4	95.7
disagree	4	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ON_TUTOR On-screen tutorials are difficult to follow.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	2	4.8	8.7	17.4
uncertain	3	4	9.5	17.4	34.8
disagree	4	15	35.7	65.2	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

LITT_UND It provided little understanding of what you are doing or how to do it.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	6	14.3	26.1	26.1
uncertain	3	5	11.9	21.7	47.8
disagree	4	12	28.6	52.2	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

OWN_PACE I felt the system allowed me to work at my own pace and direction.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	6	14.3	26.1	26.1
agree	2	17	40.5	73.9	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

PEN_PAPR I found the system was too slow because I could do the exercises much more quickly with pencil and paper.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	5	11.9	21.7	21.7
disagree	4	12	28.6	52.2	73.9
strongly disagree	5	6	14.3	26.1	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

MRE_CONT It was too quick, I would like more control of how much time I was given to solve the problems.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	1	2.4	4.3	8.7
uncertain	3	1	2.4	4.3	13.0
disagree	4	16	38.1	69.6	82.6
strongly disagree	5	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

TRY_DIFF I learnt a lot about marginal costing by trying different things with the system.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	11	26.2	47.8	47.8
uncertain	3	10	23.8	43.5	91.3
disagree	4	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

WILL_RET I am sure I will retain most of what I have learnt.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	10	23.8	43.5	43.5
uncertain	3	9	21.4	39.1	82.6
disagree	4	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ANS_PROB I learnt most from this system, by trying to answer the various problems.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	5	11.9	21.7	21.7
agree	2	14	33.3	60.9	82.6
uncertain	3	3	7.1	13.0	95.7
disagree	4	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

PROB_QUK If I use this program again, I could solve similar problems much quicker.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	5	11.9	21.7	21.7
agree	2	13	31.0	56.5	78.3
uncertain	3	5	11.9	21.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

MORE_HLP I would prefer more help from the system.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	6	14.3	26.1	26.1
agree	2	3	7.1	13.0	39.1
uncertain	3	9	21.4	39.1	78.3
disagree	4	5	11.9	21.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

FEW_HINT The help facility should have given a few hints, rather than the correct answer.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	9	21.4	39.1	43.5
uncertain	3	7	16.7	30.4	73.9
disagree	4	6	14.3	26.1	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

KEY_CORR I found it very difficult to know which keys corresponded to the command I wanted.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	3	7.1	13.0	17.4
uncertain	3	4	9.5	17.4	34.8
disagree	4	11	26.2	47.8	82.6
strongly disagree	5	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

CLER_OPT When proceeding through the program I found I was clear which options I was allowed to select.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	16	38.1	69.6	78.3
uncertain	3	2	4.8	8.7	87.0
disagree	4	2	4.8	8.7	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

GOT_LOST I hated some of the option keys, I sometimes got lost because of them.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	3	7.1	13.0	13.0
uncertain	3	4	9.5	17.4	30.4
disagree	4	11	26.2	47.8	78.3
strongly disagree	5	5	11.9	21.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

DIFF_COM When I was going through the program, pressing the same key would produce a different command.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	4.8	8.7	8.7
uncertain	3	8	19.0	34.8	43.5
disagree	4	11	26.2	47.8	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

EASYEXIT It was easy to exit from a particular route through the program.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	5	11.9	21.7	21.7
agree	2	14	33.3	60.9	82.6
uncertain	3	1	2.4	4.3	87.0
disagree	4	3	7.1	13.0	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

NOT_SKIP I couldn't skip some parts of the exercise.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	6	14.3	26.1	26.1
uncertain	3	6	14.3	26.1	52.2
disagree	4	10	23.8	43.5	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

SIM_DELT It was simple to delete any answer which I knew was wrong.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	6	14.3	26.1	26.1
agree	2	13	31.0	56.5	82.6
uncertain	3	1	2.4	4.3	87.0
disagree	4	2	4.8	8.7	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

RES_DATA I wasn't allowed to restart with the same data.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	2	4.8	8.7	17.4
uncertain	3	14	33.3	60.9	78.3
disagree	4	3	7.1	13.0	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

EASYCALC The calculator facility was easy to use.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	9	21.4	39.1	39.1
agree	2	11	26.2	47.8	87.0
uncertain	3	1	2.4	4.3	91.3
disagree	4	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

HARDCALC I found it hard to use the calculator to transfer the answer to the program.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	2.4	4.3	4.3
uncertain	3	1	2.4	4.3	8.7
disagree	4	13	31.0	56.5	65.2
strongly disagree	5	8	19.0	34.8	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

NOTCALC Using the calculator is frustrating - it doesn't do what it's supposed to do.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	2.4	4.3	4.3
uncertain	3	2	4.8	8.7	13.0
disagree	4	9	21.4	39.1	52.2
strongly disagree	5	11	26.2	47.8	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

GRAPHUSE I found the graph very useful.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	9.5	17.4	17.4
agree	2	10	23.8	43.5	60.9
uncertain	3	7	16.7	30.4	91.3
disagree	4	1	2.4	4.3	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

GRAPHDIF It was difficult to read the information from the graph.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
uncertain	3	6	14.3	26.1	30.4
disagree	4	14	33.3	60.9	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

GRAPHEAS The graphical facility allows you to compare easily the effect of different decisions on costs.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	9.5	17.4	17.4
agree	2	13	31.0	56.5	73.9
uncertain	3	6	14.3	26.1	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

MORECOLR I would have enjoyed it better if there had been more graphics, colours and sounds.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	9.5	17.4	17.4
agree	2	4	9.5	17.4	34.8
uncertain	3	8	19.0	34.8	69.6
disagree	4	7	16.7	30.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

EASY_COM The screen layout meant that it was easy to communicate with the computer.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	18	42.9	78.3	87.0
uncertain	3	3	7.1	13.0	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

LARG_INF Sometimes I was confused by the large amount of information on the screen.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	2	4.8	8.7	13.0
uncertain	3	6	14.3	26.1	39.1
disagree	4	13	31.0	56.5	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

EASY_SOL The way the problem is presented on the screen makes it easy to work out the solution.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	7.1	13.0	13.0
agree	2	14	33.3	60.9	73.9
uncertain	3	2	4.8	8.7	82.6
disagree	4	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ENGH_INF There was not enough information on the screen for me to solve the problems.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	1	2.4	4.3	13.0
uncertain	3	7	16.7	30.4	43.5
disagree	4	10	23.8	43.5	87.0
strongly disagree	5	3	7.1	13.0	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ATTR_SCR The screen was attractive and clear.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	10	23.8	43.5	52.2
uncertain	3	7	16.7	30.4	82.6
disagree	4	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

CONF_SCR The colours on the screen were very confusing.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	2.4	4.3	4.3
uncertain	3	4	9.5	17.4	21.7
disagree	4	15	35.7	65.2	87.0
strongly disagree	5	3	7.1	13.0	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

FUN_USE I really liked using this system because it was fun to use.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	10	23.8	43.5	52.2
uncertain	3	8	19.0	34.8	87.0
disagree	4	1	2.4	4.3	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

IMP_MORE I enjoyed it but it could be improved a lot more, so I would be tempted to use it again.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	7.1	13.0	13.0
agree	2	12	28.6	52.2	65.2
uncertain	3	5	11.9	21.7	87.0
disagree	4	2	4.8	8.7	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

ONE_BETT I have used a computer program like this before, but this one is a lot better.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	4	9.5	17.4	21.7
uncertain	3	12	28.6	52.2	73.9
disagree	4	5	11.9	21.7	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

SIM_READ I simply used the computer to read what is on the screen.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	4	9.5	17.4	17.4
uncertain	3	5	11.9	21.7	39.1
disagree	4	12	28.6	52.2	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

TOO_LONG The program went on too long.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	5	11.9	21.7	21.7
disagree	4	15	35.7	65.2	87.0
strongly disagree	5	3	7.1	13.0	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

EASE_USE The more you use it, the easier it becomes.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	9.5	17.4	17.4
agree	2	16	38.1	69.6	87.0
uncertain	3	2	4.8	8.7	95.7
disagree	4	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

GIVE_ASS I would be happy, if I could be given assignments using this program.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	7.1	13.0	13.0
agree	2	11	26.2	47.8	60.9
uncertain	3	5	11.9	21.7	82.6
disagree	4	2	4.8	8.7	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

NOT_LERN The program did not manage to convince me that I could learn something by using it.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
uncertain	3	4	9.5	17.4	21.7
disagree	4	14	33.3	60.9	82.6
strongly disagree	5	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

TROB_HLP I liked the way it helped me when I was in trouble.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	7.1	13.0	13.0
agree	2	12	28.6	52.2	65.2
uncertain	3	6	14.3	26.1	91.3
disagree	4	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

GOOD_REV It is a good way to revise something I have forgotten.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	5	11.9	21.7	21.7
agree	2	14	33.3	60.9	82.6
uncertain	3	4	9.5	17.4	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

NOT_INTE Most students are not interested in using such a system.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	13	31.0	56.5	56.5
disagree	4	8	19.0	34.8	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

WHY_CORR I solved the problems but I was unclear why the answer was correct.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	2.4	4.3	4.3
agree	2	1	2.4	4.3	8.7
uncertain	3	4	9.5	17.4	26.1
disagree	4	15	35.7	65.2	91.3
strongly disagree	5	2	4.8	8.7	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

MORE_EXA I would like to have more examples.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	4.8	8.7	8.7
agree	2	12	28.6	52.2	60.9
uncertain	3	4	9.5	17.4	78.3
disagree	4	4	9.5	17.4	95.7
strongly disagree	5	1	2.4	4.3	100.0
	-1	19	45.2	Missing	
		-----	-----	-----	
	Total	42	100.0	100.0	

Valid cases 23 Missing cases 19

APPENDIX D

Summaries of Responses to Open Questions

What features did you find contributed to the STRENGTH of this system?

Response	Frequency
Quick & Easy to use, lack of computer knowledge does not cause problems.	5
Lots of good examples and exercises.	5
Clarity of topics, clear and logical, good program design, main points on individual topics.	4
Self-explanatory, told me answer was wrong & how to work out it out.	3
Good pace of learning & step by step methods.	3
Simplicity.	2
Involvement in problems/exercises, allows to test knowledge.	1
Useful for specific cost accounting problems.	1
Interactive.	1
Attractive screen design.	1
Help.	1
In short time learnt basics.	1
Detailed knowledge.	1
Menu system.	1

Common responses to this open-ended question were grouped together on the basis of most frequently mentioned and of the 23 in the experimental group 2 students did not answer this question.

What features did you find contributed to the WEAKNESS of this system?

Responses	Frequency
More elaborate and self-explanatory instructions, not enough information to carry on, not user-friendly.	4
Error messages ambiguous.	3
Slow (due to floppy disk), requires too much memory.	3
Long texts/tutorials.	2
Poor feedback (more praise).	2
Problems set in a context, more incorporation examples into tutorials.	2
Should have print command, for basics like formulas.	2
More explanation about accountancy related to costs.	1
Less attractive screen.	1
Colour should be lighter.	1
Less useful graph.	1
Beeps all the time.	1
Calculator facility inhibits field selection.	1
More examples.	1
Didn't allow me to answer questions in order I wanted.	1

Common responses to this open-ended question were grouped together on the basis of most frequently mentioned and of the 23 in the experimental group 1 student did not answer this question.

If there was ONE outstanding improvement you could recommend to this system, what would it be?

Response	Frequency
Elaborate and explanatory instructions.	4
Make screen attractive, better graphics and colours, too rigid graphics.	3
More examples, keep previous exercises for review.	3
Examples in context, incorrect answer should lead to relevant information.	2
Cannot comment (due to lack of use).	2
Speed it up, run on hard disk.	2
Shorter tutorial.	1
Easier and quick access.	1
Turn off error bleep.	1
Calculator should transfer the values automatically.	1
Print out of basics.	1
Make it more interesting (less tedious).	1
Run through Windows.	1
None.	1

Common responses to this open-ended question were grouped together on the basis of most frequently mentioned and out of the 23 in the experimental group 2 students did not answer this question.

APPENDIX E

Inter-Rater Reliability of Scores

- - Correlation Coefficients - -

	NPRETSTK	NPRETSTA
NPRETSTK	1.0000 (42) P= .	0.9085 (42) P= .000
NPRETSTA	0.9085 (42) P= .000	1.0000 (42) P= .

(Coefficient / (Cases) / 1-tailed Significance)

" . " is printed if a coefficient cannot be computed

- - Correlation Coefficients - -

	POSTESTK	POSTESTA
POSTESTK	1.0000 (42) P= .	0.9445 (42) P= .000
POSTESTA	0.9445 (42) P= .000	1.0000 (42) P= .

(Coefficient / (Cases) / 1-tailed Significance)

" . " is printed if a coefficient cannot be computed

APPENDIX F

Comparison of experimental and control differences using Independent T-tests

.

t-tests for Independent Samples of GROUP expt/contrl

Variable IMPROVEM	Number of Cases	Mean	SD	SE of Mean
GROUP >= 4	23	-3.3587	3.230	0.674
GROUP <= 4	19	-1.0789	2.328	0.534

Mean Difference = -2.2797

Levene's Test for Equality of Variances: F= 5.253 P= .027

t-test for Equality of Means

Variances	t-value	df	2-Tail Sig	SE of Diff	95% CI for Diff
Equal	-2.57	40	0.014	0.887	(-4.072, -0.488)
Unequal	-2.65	39.35	0.011	0.860	(-4.018, -0.542)

t-tests for Independent Samples of GROUP expt/contrl

Variable IMPROVEM	Number of Cases	Mean	SD	SE of Mean
GROUP 1	9	-1.5000	2.788	0.929
GROUP 4	12	-2.9688	3.250	0.938

Mean Difference = 1.4688

Levene's Test for Equality of Variances: F= .505 P= .486

t-test for Equality of Means

Variances	t-value	df	2-Tail Sig	SE of Diff	95% CI for Diff
Equal	1.09	19	0.291	1.351	(-1.359, 4.297)
Unequal	1.11	18.58	0.280	1.321	(-1.300, 4.237)

t-tests for Independent Samples of GROUP expt/contrl

Variable IMPROVEM	Number of Cases	Mean	SD	SE of Mean
GROUP 2	4	-0.9688	2.938	1.469
GROUP 5	5	-5.7500	3.300	1.476

Mean Difference = 4.7813

Levene's Test for Equality of Variances: F= .370 P= .562

t-test for Equality of Means

Variances	t-value	df	2-Tail Sig	SE of Diff	95% CI for Diff
Equal	2.26	7	0.058	2.113	(-0.215, 9.778)
Unequal	2.30	6.87	0.056	2.082	(-0.162, 9.724)

t-tests for Independent Samples of GROUP expt/contrl

Variable IMPROVEM	Number of Cases	Mean	SD	SE of Mean
GROUP 3	6	-0.5208	1.091	0.445
GROUP 6	6	-2.1458	2.472	1.009

Mean Difference = 1.6250

Levene's Test for Equality of Variances: F= 8.836 P= .014

t-test for Equality of Means

Variances	t-value	df	2-Tail Sig	SE of Diff	95% CI for Diff
Equal	1.47	10	0.171	1.103	(-0.833, 4.083)
Unequal	1.47	6.88	0.185	1.103	(-0.993, 4.243)

APPENDIX G

Reliability of the Attitudinal Scales using Cronbach's Alpha Coefficient of Reliability

COMPUTING KNOWLEDGE - COMPUT

RELIABILITY ANALYSIS - SCALE (ALPHA)

1.	PRIOR_KN	prior comp. required
2.	ZNOTLEAR	didn't learn anything
3.	CONFIDEN	confident @ learn
4.	ZFIGHT	fight the computer
5.	OUT_MANU	well without manuals
6.	ZWHAT_TD	not see comp. want

Reliability Coefficients

N of Cases = 23.0

N of Items = 6

Alpha = .4785

COMPARISON WITH OTHER MEDIA - COMPAR

RELIABILITY ANALYSIS - SCALE (ALPHA)

1.	BET_TEAC	comp. better teachers
2.	ZLEC_CLO	lecturer close by
3.	ONE_ONE	one-to-one tutor
4.	ZUNANSW_	lots of questions
5.	ALL_STEP	didn't miss steps
6.	CLASSTUT	better than tutorial
7.	ZBOOK	prefer accou. book
8.	I_ACTIVE	prefer due interactive
9.	REL_VAR	find relation variab.
10.	ZPEN_PAP	prefer pencil & paper
11.	SEE_HOW	can see how it done

Reliability Coefficients

N of Cases = 23.0

N of Items = 11

Alpha = .8389

INDIVIDUAL FLEXIBILITY - INDIVID

RELIABILITY ANALYSIS - SCALE (ALPHA)

1.	ZADV_ACC	for advance. students
2.	COM_HARD	comp. got harder
3.	ZLONGSTE	comp. labor. steps
4.	IND_NEDS	comp. adjusts needs
5.	JUM_DEMD	more demand. stuff

Reliability Coefficients

N of Cases = 23.0

N of Items = 5

Alpha = .0871

ACCOUNTANCY KNOWLEDGE/TEACHING - ACC

RELIABILITY ANALYSIS - SCALE (ALPHA)

1.	PREV_ACC	no know. of accoun.
2.	GOOD_INT	good intro marg cost
3.	ZONLY_AC	use familiar accoun.
4.	BECM_EXP	i became an expert
5.	ZDIFF_BA	diff. to learn basics
6.	GROUNDIN	1st class grounding
7.	ZWORK_OU	hard to work prob.
8.	FULLTEST	fully test understand.
9.	ZNO_THIN	didn't know marginal
10.	KN_INCRE	acc. know. increased
11.	EASY_KNO	easily use my know.
12.	TEACELSE	use teach someone
13.	GUID_ME	comp. guide. learning
14.	OTHER_TO	learn other topics
15.	ACCU_PIC	accurate marginal pic
16.	RELA_BEH	relation. behind marg
17.	ZNOT_REL	don't under. relation.

Reliability Coefficients

N of Cases = 23.0

N of Items = 17

Alpha = .7994

ERROR MESSAGES AND DOCUMENTATION - ERROR

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

1.	HELPSTUC	mess. help got stuck
2.	ZMESS_HE	messages unhelpful
3.	MAK_MIST	allows me mistakes
4.	ZNOT_COR	not allow incorrect values
5.	ZPROG_UN	prog. so unforgiving
6.	ZERR_MS	error message for incorrect value
7.	PREC_WRG	told precisely wrong

Reliability Coefficients

N of Cases = 23.0

N of Items = 7

Alpha = .5714

EXPLANATION (FEEDBACK) INSTRUCTIONS - EXPLAN

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

1.	ZGOOD_FE	only good feedback
2.	ZENGH_PR	not enough praise
3.	HELP_STP	help me every step
4.	IMM_FEED	immediate feedback
5.	ZEXPL_UN	unclear explanations
6.	CLE_EXPL	clear explanations
7.	IND_FEED	individual feedback to each student
8.	ZAİM_EXE	prefer aim exercise

Reliability Coefficients

N of Cases = 23.0

N of Items = 8

Alpha = .1808

ROUTE THROUGH PROGRAM - ROUTE

RELIABILITY ANALYSIS - SCALE (ALPHA)

1.	ZALLMEAN	didn't know all menu items meant
2.	ZINST_CL	instructions unclear how to proceed
3.	SAME_MIS	don't same mistakes
4.	ZHARD_TR	hard to keep track which bits done
5.	ANY_ORD	problems any order
6.	ZDIFF_RO	unable different route
7.	ZSUGG_RO	prefer suggest. route
8.	ZPROM_CL	prompts unclear from screen
9.	ZRET_LEF	prefer return I left off
10.	ZON_TUTO	tutorials unclear from screen
11.	ZLITT_UN	little understanding of what doing

Reliability Coefficients

N of Cases = 23.0

N of Items = 11

Alpha = .6575

CONTROL OF LEARNING - CONT

RELIABILITY ANALYSIS - SCALE (ALPHA)

1.	OWN_PACE	my own pace and direction
2.	ZPEN_PAR	prefer pencil & paper
3.	ZMRE_CON	prefer more control of time
4.	TRY_DIFF	learn try different things with system

Reliability Coefficients

N of Cases = 23.0

N of Items = 4

Alpha = .0240

RETENTION AND LEARNING - RETEN

RELIABILITY ANALYSIS - SCALE (ALPHA)

1.	WILL_RET	I will retain most of what I learnt
2.	ANS_PROB	learnt lots problems
3.	PROB_QUK	solve problem quick use this again

Reliability Coefficients

N of Cases = 23.0

N of Items = 3

Alpha = .5149

HELP FACILITY - HELP

RELIABILITY ANALYSIS - SCALE (ALPHA)

1.	ZMORE_HL	prefer more help
2.	ZFEW_HIN	prefer few hints

Reliability Coefficients

N of Cases = 23.0

N of Items = 2

Alpha = .6834

FUNCTIONS OR KEYS - FUNCTIONS

RELIABILITY ANALYSIS - SCALE (ALPHA)

1.	ZKEY_COR	which keys corresponded
2.	CLER_OPT	clear which options allow select
3.	ZGOT_LOS	got lost because of options keys
4.	ZDIFF_CO	different command - same key!
5.	EASYEXIT	It was easy to exit
6.	ZNOT_SKI	couldn't skip some parts of exercise
7.	SIM_DELT	simple to delete any answer
8.	ZRES_DAT	not restart same data

Reliability Coefficients

N of Cases = 23.0

N of Items = 8

Alpha = .5030

CALCULATOR FACILITY - CALC

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

1.	EASYCALC	calculator easy to use
2.	ZHARDCAL	complex calculator - hard to use
3.	ZNOTCAL	calculator does not do what you want

Reliability Coefficients

N of Cases = 23.0 N of Items = 3

Alpha = .8571

GRAPHICS FACILITY - GRAPH

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

1.	GRAPHUSE	graph very useful
2.	ZGRAPHDI	difficult to read graph
3.	GRAPHEAS	easy compar. graph
4.	ZMORECOL	more graphics, colours & sounds

Reliability Coefficients

N of Cases = 23.0 N of Items = 4

Alpha = .1789

PRESENTATION INFO. ON SCREEN - PRESENT

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

1.	EASY_COM	easy communicate with computer
2.	ZLARG_IN	confus. large info. on the screen
3.	EASY_SOL	easy work out solution
4.	ZENGH_IN	not much info. on the screen
5.	ATTR_SCR	attractive screen and clear
6.	ZCONF_SC	colour very confusing on screen

Reliability Coefficients

N of Cases = 23.0 N of Items = 6

Alpha = .6261

HOLISTIC SYSTEM - HOL

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

1.	FUN_USE	comp. fun to use
2.	ZIMP_MOR	prefer more improve.
3.	ONE_BETT	comp. much better than others
4.	ZSIM_REA	simply read screen
5.	ZTOO_LON	prog. went too long
6.	EASE_USE	more use more easy
7.	GIVE_ASS	like assignments to be given
8.	ZNOT_LER	didn't learn anything
9.	TROB_HLP	helped me I trouble
10.	GOOD_REV	way to revise something forgotten

Reliability Coefficients

N of Cases = 23.0

N of Items = 10

Alpha = .6142

OVERALL PERCEPTION - PERCEP

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

1.	ZNOT_INT	student not use system
2.	ZWHY_COR	unclear why correct
3.	ZMORE_EX	prefer more example

Reliability Coefficients

N of Cases = 23.0

N of Items = 3

Alpha = .4216

APPENDIX H

Factor Analysis of Accountancy Knowledge/Teaching, Calculator facility and Comparison with other Media

Analysis number 1 Listwise deletion of cases with missing values

Correlation Matrix:

	ACC	CALC	COMPAR
ACC	1.00000		
CALC	-0.14949	1.00000	
COMPAR	0.56287	-0.24440	1.00000

Determinant of Correlation Matrix = 0.6422281

Kaiser-Meyer-Olkin Measure of Sampling Adequacy = 0.53983

Bartlett Test of Sphericity = 8.93004, Significance = 0.03024

Extraction 1 for analysis 1, Principal Components Analysis (PC)

Initial Statistics:

Variable	Communality	* *	Factor	Eigenvalue	Pct of Var	Cum Pct
ACC	1.00000	*	1	1.67792	55.9	55.9
CALC	1.00000	*	2	0.89390	29.8	85.7
COMPAR	1.00000	*	3	0.42818	14.3	100.0

PC extracted 1 factors.

Factor Matrix:

	Factor 1
ACC	0.82748
CALC	-0.49443
COMPAR	0.86530

Final Statistics:

Variable	Communality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
		*				
ACC	0.68472	*	1	1.67792	55.9	55.9
CALC	0.24446	*				
COMPAR	0.74874	*				

Reproduced Correlation Matrix:

	ACC	CALC	COMPAR
ACC	0.68472*	0.25964	-0.15315
CALC	-0.40913	0.24446*	0.18343
COMPAR	0.71602	-0.42783	0.74874*

The lower left triangle contains the reproduced correlation matrix; the diagonal, reproduced communalities; and the upper right triangle residuals between the observed correlations and the reproduced correlations.

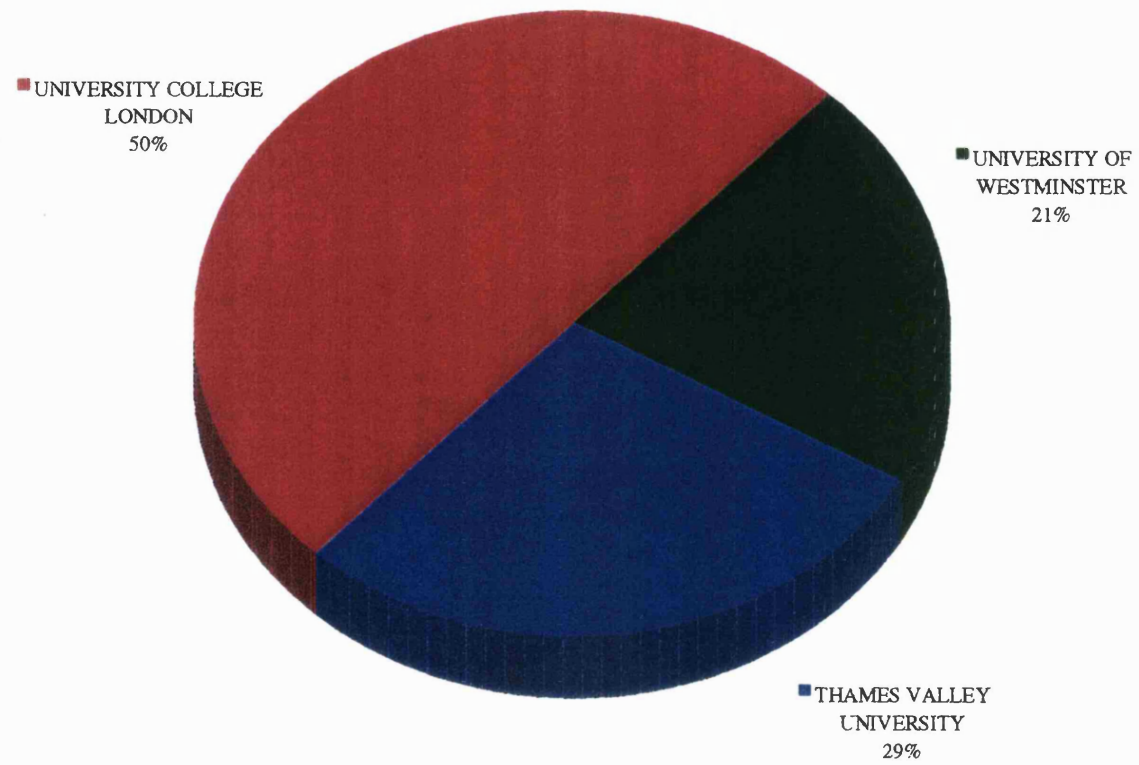
There are 3 (100%) residuals (above diagonal) with absolute values > 0.05.

Skipping rotation 1 for extraction 1 in analysis 1

APPENDIX I

Percentage of participants from the different universities

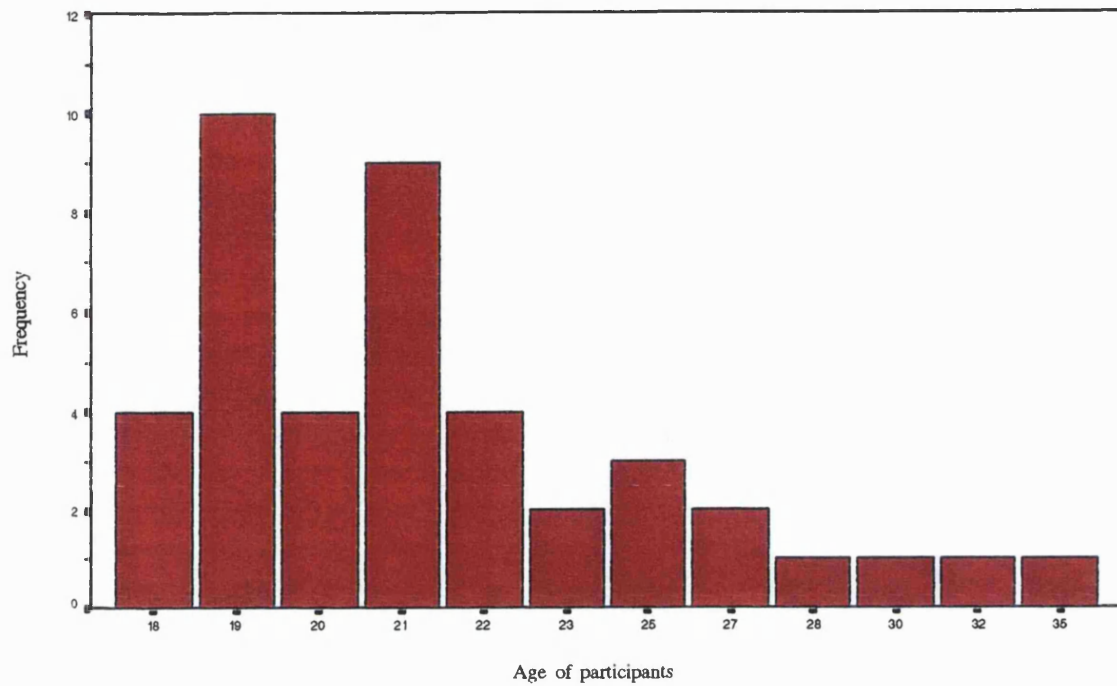
Percentage of participants from the different universities



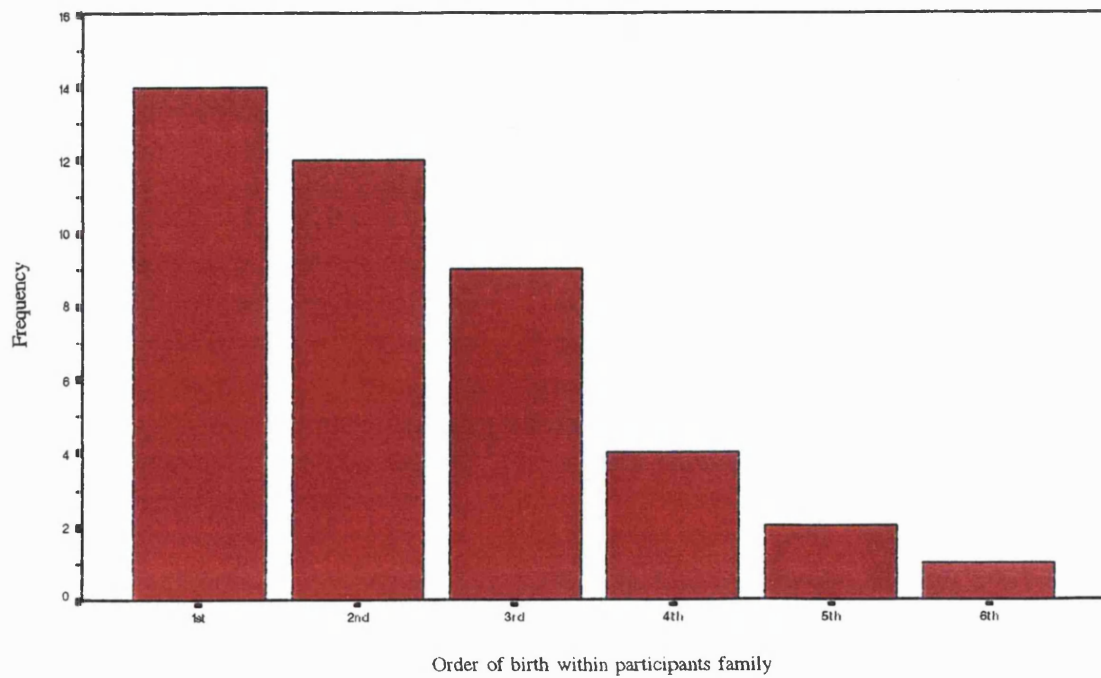
APPENDIX J

**Age distribution of participants (in years)
and Order of birth distribution of participants**

Age distribution of participants (in years)



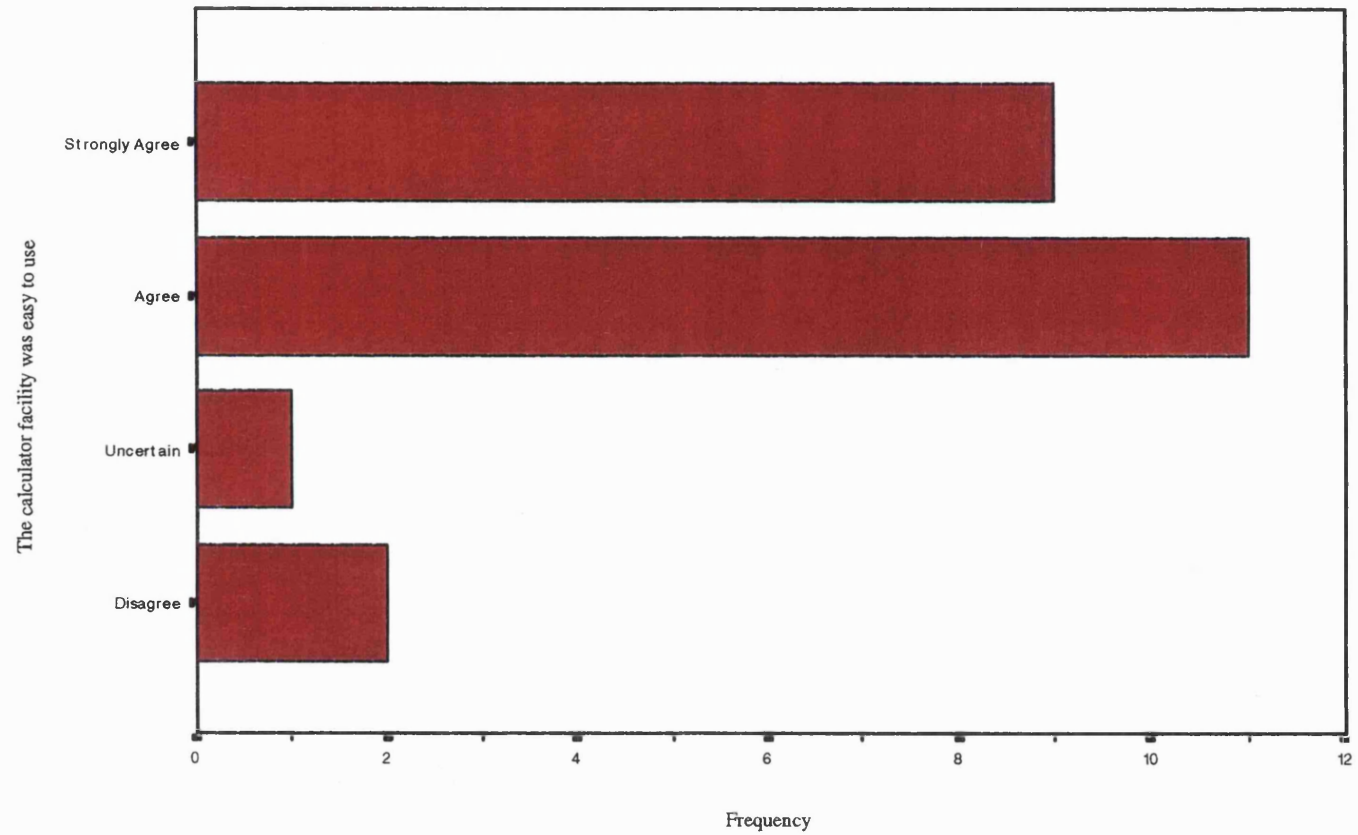
Order of birth distribution of participants



APPENDIX K

Responses to "The calculator facility was easy to use"

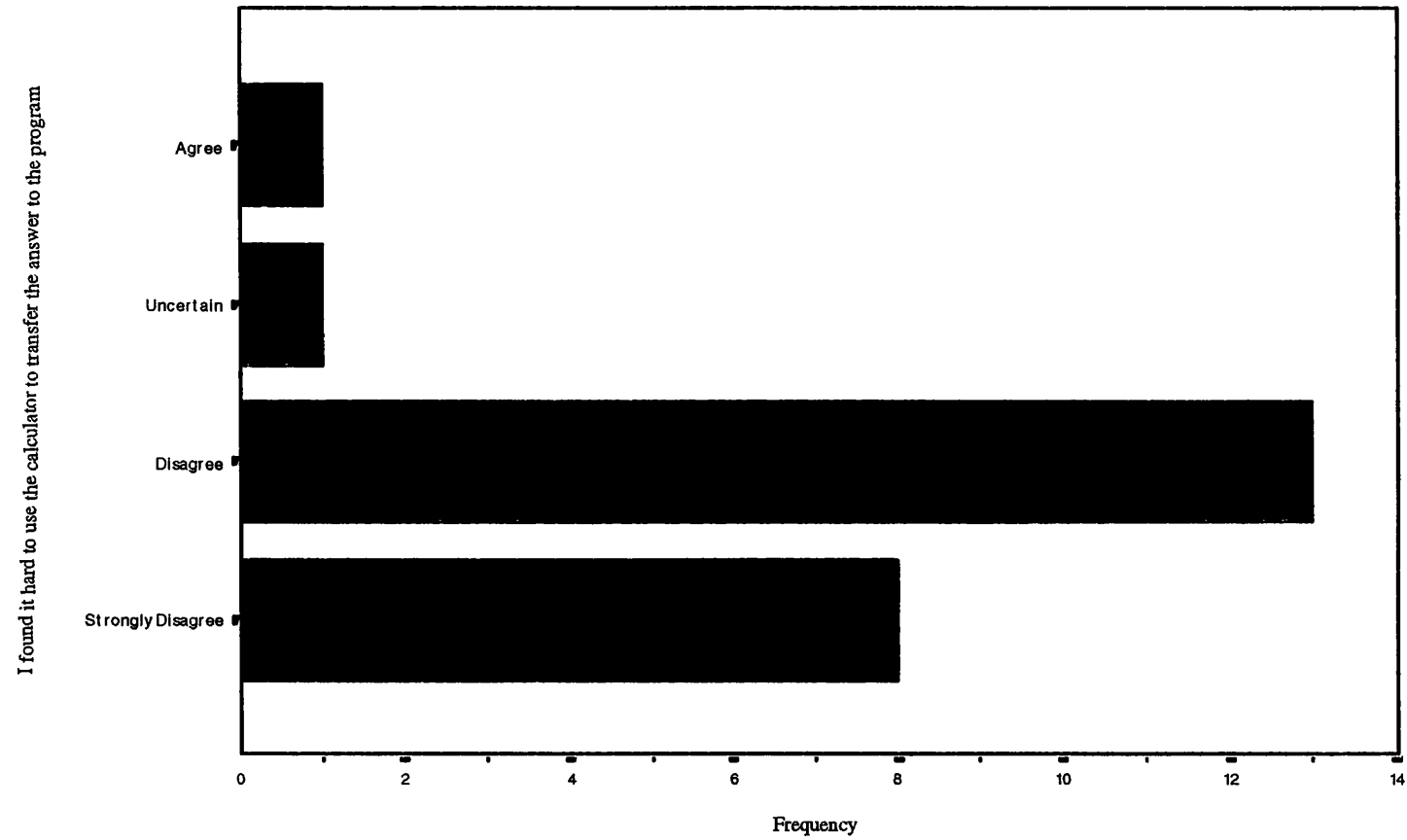
Responses to “The calculator facility was easy to use”



APPENDIX L

**Responses to "I found it hard to use the calculator
to transfer the answer to the program"**

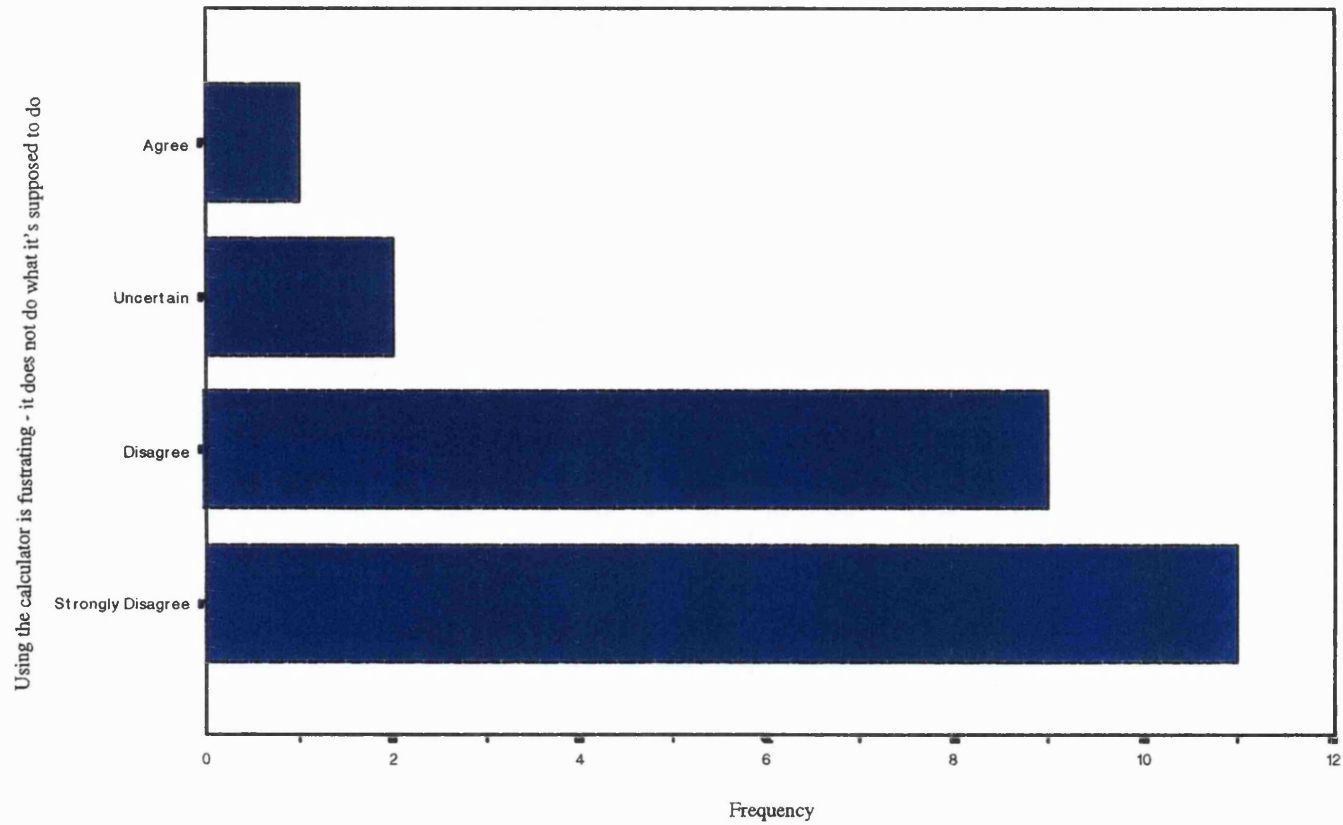
Responses to “ I found it hard to use the calculator to transfer the answer to the program”



APPENDIX M

**Responses to "Using the calculator is frustrating
- it does not do what it's supposed to do "**

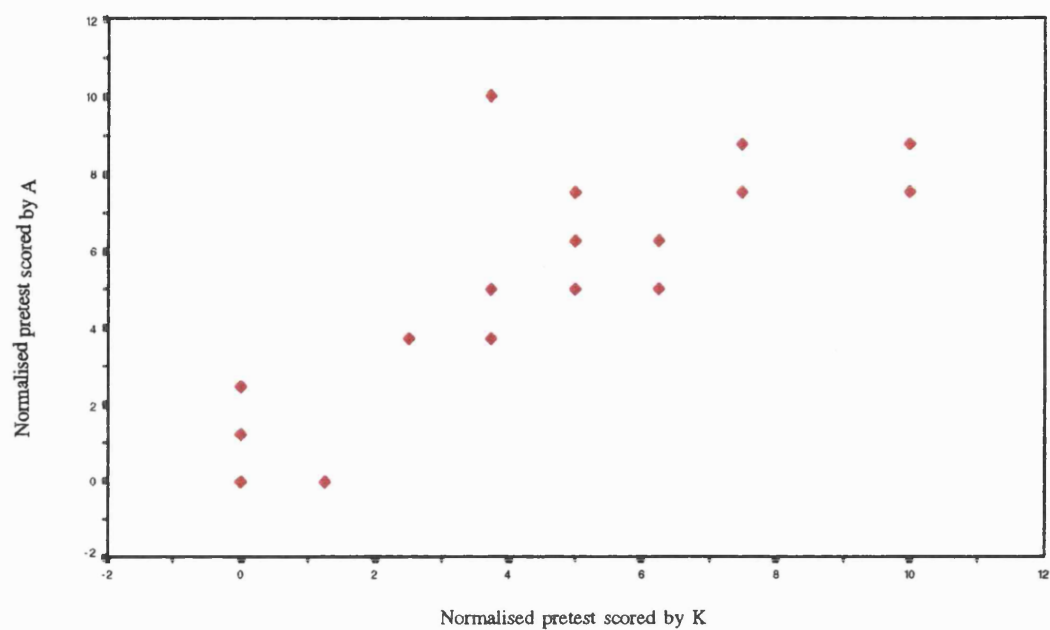
Responses to “Using the calculator is frustrating - it does not do what it’s supposed to do”



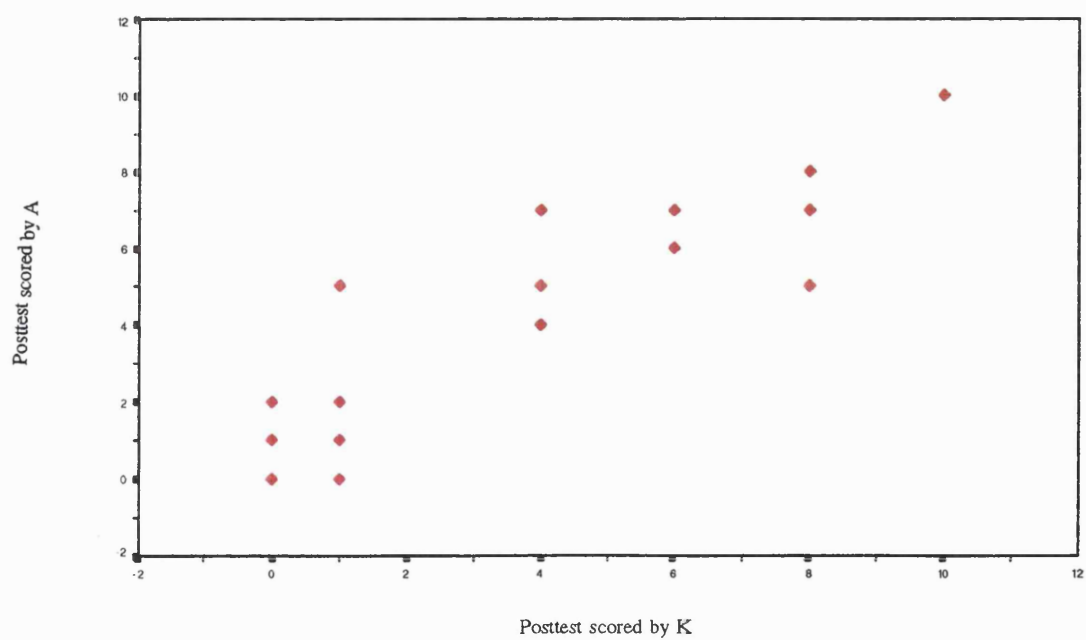
APPENDIX N

**Scattergram of pretest scores marked by two raters A and K
and Scattergram of posttest scores marked by two raters A and K**

Scattergram of pretest scores marked by two raters A and K



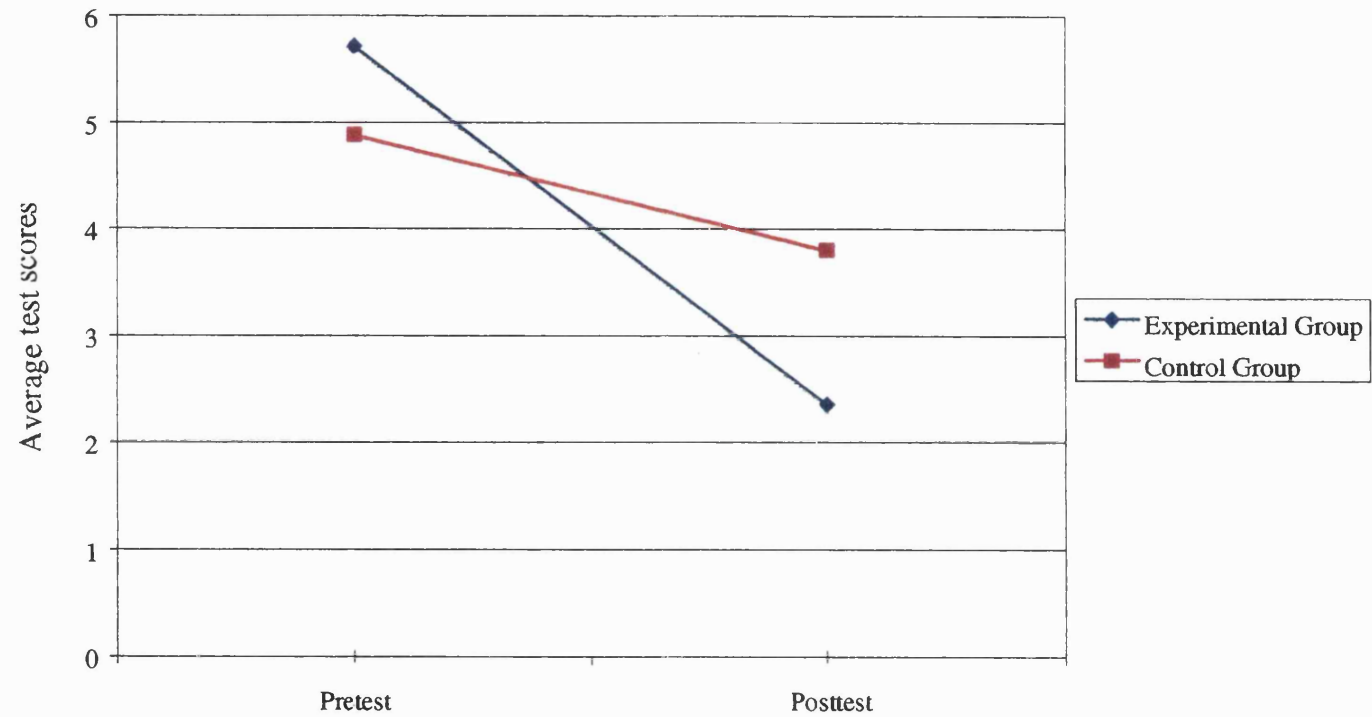
Scattergram of posttest scores marked by two raters A and K



APPENDIX O

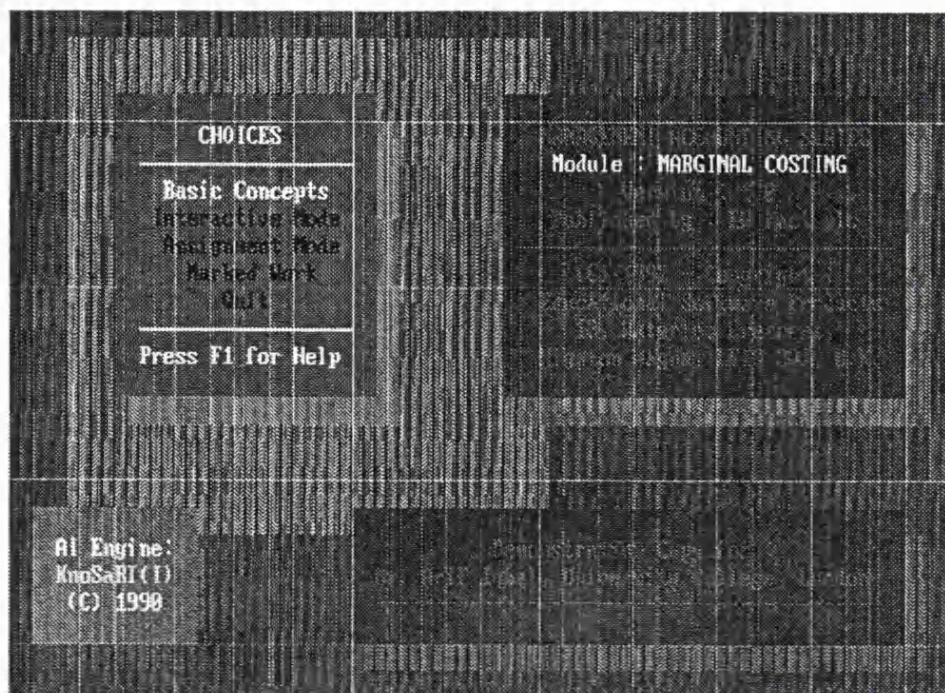
Average group scores before and after Computer or Classroom Instruction

Average group scores before and after Computer or Classroom Instruction



APPENDIX P

Screen Dumps



T6 - Profits and Margin of safety	
We can calculate the margin of safety by first calculating the break even sales and then deducting it from the target revenue.	
$B/E \text{ Sales} = \text{Fixed costs} / C/S \text{ ratio} \quad \text{Revenue required to generate a contribution equal to fixed costs}$	
	Co. B
Fixed costs	4,000
C/S ratio	30%
B/E Sales	13,333
Target revenue	20,000
M/S Sales	6,667
M/S Sales	6,667
Alt N : Next Topic	Alt P : Previous Topic
F1 Help	Esc to Exit
t, l, PgUp, PgDn to scroll	
Alt I Index of Topics	
FB Calculator	

T1	T6 - Profits and Margin of safety	
Basic uses of Marginal Costing :		
		Topic Title
(a) quick comparisons between alternatives		
(b) extra order profitability calculator	T1	Variable & fixed costs
(c) break even calculations	T2	Contribution & C/S ratio
(d) MARGIN OF SAFETY CALCULATIONS	T3	Profitability of Alternatives
(e) make or buy decisions	T4	Profitability of Extra orders
	T5	Profitability & Breaking even
	T6	Profits and Margin of safety
	T7	Profitability - Make or buy
MARGIN OF SAFETY CALCULATIONS :		
The balance of the make up of total cost is the last element in strategic choice. For example, consider the following to highlight choice & press Enter		
Revenue		
	20,000	
Alt N : Next Topic	Alt P : Previous Topic	Alt I : Index of Topics
F1 Help	Esc to Exit	F8 Calculator
T1, T4, PgUp, PgDn to scroll		

Computer generated example				
Quantity ?	Sale Price	Revenue	Calc - Output	
	102.4800	Total VC		
	65.5200	Total Cost	C/S Ratio %	
		Fixed Cost	B/E Units	B/E £ Sales
		Profit	M/S Units	M/S £ Sales
B/E : Break Even			48476.14	
M/S : Margin of Safety				
Instructions		Interactive messages		
Please use ↑ or ↓ to move to the required box and enter a value in it. Press Esc to exit.				
F1 Help	F3 Erase entries	F5 What-if	F6 Graph	F8 Calculator
1F8 Save				
(C) Educational Software Products(1992)				

Computer generated example				
Quantity ?	Sale Price	Revenue	Calc Output	
			24000.0000	
	UC per unit 102.4800	Total UC		
	Cont./Unit 65.5200	Total Cont.	C/S Ratio %	
		4054902		
		Fixed Cost	B/E Units	B/E I Sales
		Profit	M/S Units	M/S I Sales
			48476.14	

24000

Calculator

1	2	3	*
4	5	6	-
7	8	9	+
0	.	=	/

Esc to quit

Interactive messages

Pressing = or Enter again will transfer the result to Quantity

F1 Help F3 Erase entries F5 What-if F6 Graph F8 Calculator F10 Save

(C) Educational Software Products(1992)

Computer generated example				
Quantity 57374	Sale Price 25.0000	Revenue 1434350	Calc Output	
			24000.0000	
	UC per unit 6.7500	Total UC 387275		
	Cont./Unit 18.2500	Total Cont. 1047076	C/S Ratio % 73.00	
		Fixed Cost 137163	B/E Units 7515.78	B/E I Sales 187895
		Profit 909913	M/S Units 49858.25	M/S I Sales 1246456

Instructions

Please use ↑ or ↓ to move to the required box and enter a value in it.
Press Esc to exit.

Interactive messages

F1 Help F3 Erase entries F5 What-if F6 Graph F8 Calculator F10 Save

(C) Educational Software Products(1992)

Computer generated example		
Example entry and solution		
A marginal costing decision involves the fourteen variables shown on the screen. They have been laid out in an easy to follow layout to make some of the inter-relationships instinctively obvious. Some of these variables will be given, as a matter of fact. Some will be based on the manager's decisions. Together, these two types of variables constitute the independent variables. All the remaining variables are dependent. They are derived from the independent variables : You cannot change them till the constraining independent variables are changed.		
In this part of the program, you can either load the independent variables from a disk, enter them yourself (subject to constraints in case of interactive mode) or let the computer pick at random what variables will be treated as independent and generate random values for them. You have to calculate and fill in the remaining values. In case of interactive mode, you will get immediate feedback and		
Esc to Exit	F1,F4,FgUp,FgDn to scroll	F8 Calculator
F1	(C) Educational Software Products(1992)	

Computer generated example					
Quantity	Sale Price	Revenue	Cost	Profit	
	173.8808				
	UC per unit	Total UC			
	Cont./Unit	Total Cont	C/S Ratio %		
		2498895	32.00		
		Fixed Cost	B/E Units	B/E £ Sales	
		686141		2144191	
		Profit	M/S Units	M/S £ Sales	
B/E : Break Even		?			
M/S : Margin of Safety					
Instructions		Interactive messages			
Please use ↑ or ↓ to move to the required box and enter a value in it. Press Esc to exit.		Sorry! That isn't the correct answer. Please try again ! Suggested relationship : Profit = Total Cont. - Fixed Cost			
F1 Help	F3 Erase entries	F5 What-if	F6 Graph	F8 Calculator	F10 Save
(C) Educational Software Products(1992)					

APPENDIX Q

APT Frequency Distribution of Questionnaire Responses

UNIV university of subject

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	dmu	12	60.0	60.0	60.0
	nwi	8	40.0	40.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

AGE age of subject

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	18	3	15.0	15.0	15.0
	19	4	20.0	20.0	35.0
	20	1	5.0	5.0	40.0
	21	2	10.0	10.0	50.0
	22	1	5.0	5.0	55.0
	25	1	5.0	5.0	60.0
	26	1	5.0	5.0	65.0
	27	2	10.0	10.0	75.0
	28	1	5.0	5.0	80.0
	29	1	5.0	5.0	85.0
	33	1	5.0	5.0	90.0
	38	1	5.0	5.0	95.0
	39	1	5.0	5.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

OCCUP occupation of subject

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
UNDERGRADUATE	1	12	60.0	60.0	60.0
OTHER	3	8	40.0	40.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

SEX gender of subject

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
male	1	9	45.0	45.0	45.0
female	2	11	55.0	55.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

BIRTH_OR order of birth within subjects family

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
1st	1	7	35.0	35.0	35.0
2nd	2	6	30.0	30.0	65.0
3rd	3	4	20.0	20.0	85.0
4th	4	2	10.0	10.0	95.0
5th	5	1	5.0	5.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

HOME_COM computer at home?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	15	75.0	75.0	75.0
no	2	5	25.0	25.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

HOME_USE use home computer?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	14	70.0	70.0	70.0
no	2	1	5.0	5.0	75.0
not applicable	3	5	25.0	25.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

HOME_OPE operate home computer?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	13	65.0	68.4	68.4
no	2	1	5.0	5.3	73.7
not applicable	3	5	25.0	26.3	100.0
	-1	1	5.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 19 Missing cases 1

COLL_COM computer at college/work?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	19	95.0	95.0	95.0
no	2	1	5.0	5.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

COLL_USE use college computer?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	17	85.0	85.0	85.0
no	2	2	10.0	10.0	95.0
not applicable	3	1	5.0	5.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

COLL_OPE operate college computer?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	16	80.0	84.2	84.2
no	2	2	10.0	10.5	94.7
not applicable	3	1	5.0	5.3	100.0
	-1	1	5.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 19 Missing cases 1

ENJOY_CO do you enjoy using computer?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
not much	3	4	20.0	20.0	20.0
unsure	4	5	25.0	25.0	45.0
quite a lot	5	11	55.0	55.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

COMP_SCH how often use school computer?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
never	1	3	15.0	15.0	15.0
rarely	2	7	35.0	35.0	50.0
monthly	3	5	25.0	25.0	75.0
weekly	4	5	25.0	25.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

COMP_EDU any computer education/training?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	17	85.0	85.0	85.0
no	2	3	15.0	15.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

TYPE_EDU types of education/training?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
not applicable	1	2	10.0	10.0	10.0
Databases	7	1	5.0	5.0	15.0
Spreadsheets	8	2	10.0	10.0	25.0
Word processing	9	1	5.0	5.0	30.0
N/A, Spread	18	2	10.0	10.0	40.0
Spread, Word	89	1	5.0	5.0	45.0
Data, Spead, Word	789	4	20.0	20.0	65.0
O Level, Data, Spread,	2789	5	25.0	25.0	90.0
Data, Spread, Word, O	7890	2	10.0	10.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

COMP_LAN program in any comp. languages?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
yes	1	3	15.0	15.0	15.0
no	2	17	85.0	85.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

NO_COMP no of comp. langs. ?

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
not applicable	0	17	85.0	85.0	85.0
one language	1	2	10.0	10.0	95.0
two languages	2	1	5.0	5.0	100.0
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

PRIOR_KN prior comp. required

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	15.0	30.0	30.0
agree	2	7	35.0	70.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

NOTLEARN didn't learn anything

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
disagree	4	4	20.0	40.0	50.0
strongly disagree	5	5	25.0	50.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

CONFIDEN confident @ learn

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
agree	2	7	35.0	70.0	80.0
uncertain	3	1	5.0	10.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

FIGHT fight the computer

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
disagree	4	8	40.0	80.0	80.0
strongly disagree	5	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

OUT_MANU well without manuals

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	15.0	30.0	30.0
agree	2	6	30.0	60.0	90.0
strongly disagree	5	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

WHAT_TDO not see comp. want

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	1	5.0	10.0	10.0
disagree	4	5	25.0	50.0	60.0
strongly disagree	5	4	20.0	40.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

BET_TEAC comp. better teachers

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
uncertain	3	3	15.0	30.0	40.0
disagree	4	3	15.0	30.0	70.0
strongly disagree	5	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

LEC_CLOS lecturer close by

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	5	25.0	50.0	70.0
uncertain	3	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

ONE_ONE one-to-one tutor

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	7	35.0	70.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

UNANSW_Q lots of questions

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	3	15.0	30.0	30.0
disagree	4	5	25.0	50.0	80.0
strongly disagree	5	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

ALL_STEP didn't miss steps

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	6	30.0	60.0	80.0
uncertain	3	1	5.0	10.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

CLASSTUT better than tutorial

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	20.0	40.0	40.0
agree	2	5	25.0	50.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

BOOK prefer accou. book

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	1	5.0	10.0	10.0
disagree	4	6	30.0	60.0	70.0
strongly disagree	5	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

I_ACTIVE prefer due interactive

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	20.0	40.0	40.0
agree	2	5	25.0	50.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

REL_VAR find relation variab.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	5	25.0	50.0	50.0
uncertain	3	3	15.0	30.0	80.0
disagree	4	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

PEN_PAP prefer pencil & paper

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	1	5.0	10.0	10.0
disagree	4	6	30.0	60.0	70.0
strongly disagree	5	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

SEE_HOW can see how it done

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	5	25.0	50.0	50.0
uncertain	3	3	15.0	30.0	80.0
disagree	4	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

ADV_ACC for advance. students

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	5.0	10.0	10.0
uncertain	3	4	20.0	40.0	50.0
disagree	4	5	25.0	50.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

COM_HARD comp. got harder

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	3	15.0	30.0	30.0
disagree	4	5	25.0	50.0	80.0
strongly disagree	5	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

LONGSTEP comp. labor. steps

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
agree	2	3	15.0	30.0	40.0
uncertain	3	4	20.0	40.0	80.0
disagree	4	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

IND_NEDS comp. adjusts needs

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	3	15.0	30.0	30.0
disagree	4	4	20.0	40.0	70.0
strongly disagree	5	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

JUM_DEMD more demand. stuff

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	4	20.0	40.0	40.0
uncertain	3	3	15.0	30.0	70.0
disagree	4	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

PREV_ACC no know. of accoun.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	3	15.0	30.0	30.0
uncertain	3	4	20.0	40.0	70.0
disagree	4	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

GOOD_INT good intro marg cost

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	4	20.0	40.0	40.0
uncertain	3	4	20.0	40.0	80.0
disagree	4	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

ONLY_ACC use familiar accoun.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	10.0	20.0	20.0
uncertain	3	5	25.0	50.0	70.0
disagree	4	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

BECM_EXP i became an expert

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	10.0	20.0	20.0
uncertain	3	6	30.0	60.0	80.0
disagree	4	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

DIFF_BAS diff. to learn basics

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	5.0	10.0	10.0
uncertain	3	4	20.0	40.0	50.0
disagree	4	5	25.0	50.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

GROUNDIN 1st class grounding

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	6	30.0	60.0	60.0
uncertain	3	2	10.0	20.0	80.0
disagree	4	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

WORK_OUT hard to work prob.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	1	5.0	10.0	10.0
disagree	4	6	30.0	60.0	70.0
strongly disagree	5	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

FULLTEST fully test understand.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	4	20.0	40.0	40.0
uncertain	3	5	25.0	50.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

NO_THING didn't know marginal

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	5.0	10.0	10.0
uncertain	3	1	5.0	10.0	20.0
disagree	4	5	25.0	50.0	70.0
strongly disagree	5	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

KN_INCRE acc. know. increased

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	15.0	30.0	30.0
agree	2	3	15.0	30.0	60.0
uncertain	3	3	15.0	30.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

EASY_KNO easily use my know.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
agree	2	3	15.0	30.0	40.0
uncertain	3	5	25.0	50.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

TEACELSE use teach someone

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
agree	2	6	30.0	60.0	70.0
uncertain	3	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

GUID_ME comp. guide. learning

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	6	30.0	60.0	60.0
uncertain	3	3	15.0	30.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

OTHER_TO learn other topics

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	5	25.0	50.0	50.0
agree	2	4	20.0	40.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

ACCU_PIC accurate marginal pic

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	20.0	40.0	40.0
agree	2	5	25.0	50.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

RELA_BEH relation. behind marg

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	4	20.0	40.0	40.0
uncertain	3	4	20.0	40.0	80.0
disagree	4	1	5.0	10.0	90.0
strongly disagree	5	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

NOT_RELA don't under. relation.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
agree	2	2	10.0	20.0	30.0
uncertain	3	5	25.0	50.0	80.0
disagree	4	1	5.0	10.0	90.0
strongly disagree	5	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

HELPSTUC mess. help got stuck

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	3	15.0	30.0	30.0
uncertain	3	2	10.0	20.0	50.0
disagree	4	4	20.0	40.0	90.0
strongly disagree	5	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

MESS_HEL messages unhelpful

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	2	10.0	20.0	40.0
uncertain	3	1	5.0	10.0	50.0
disagree	4	3	15.0	30.0	80.0
strongly disagree	5	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

MAK_MIST allows me mistakes

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	3	15.0	30.0	50.0
uncertain	3	3	15.0	30.0	80.0
disagree	4	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

NOT_CORR not allow incorrect values

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	3	15.0	30.0	30.0
uncertain	3	1	5.0	10.0	40.0
disagree	4	5	25.0	50.0	90.0
strongly disagree	5	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

PROG_UNF prog. so unforgiving

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	15.0	30.0	30.0
agree	2	5	25.0	50.0	80.0
uncertain	3	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

ERR_MSG error message for incorrect value

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	3	15.0	30.0	30.0
uncertain	3	6	30.0	60.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

PREC_WRG told precisely wrong

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	5	25.0	50.0	50.0
agree	2	4	20.0	40.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

GOOD_FED only good feedback

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	7	35.0	70.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

ENGH_PRA not enough praise

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	15.0	30.0	30.0
agree	2	4	20.0	40.0	70.0
uncertain	3	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

HELP_STP help me every step

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	7	35.0	70.0	70.0
uncertain	3	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

IMM_FEED immediate feedback

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	6	30.0	60.0	80.0
uncertain	3	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

EXPL_UND unclear explanations

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	1	5.0	10.0	10.0
disagree	4	7	35.0	70.0	80.0
strongly disagree	5	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

CLE_EXPL clear explanations

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	20.0	40.0	40.0
agree	2	5	25.0	50.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

IND_FEED individual feedback to each student

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	3	15.0	30.0	30.0
uncertain	3	7	35.0	70.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

AIM_EXER prefer aim exercise

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	4	20.0	40.0	40.0
disagree	4	2	10.0	20.0	60.0
strongly disagree	5	4	20.0	40.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

ALLMEANT didn't know all menu items meant

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	3	15.0	30.0	30.0
uncertain	3	2	10.0	20.0	50.0
disagree	4	5	25.0	50.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

INST_CLR instructions unclear how to proceed

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
agree	2	5	25.0	50.0	60.0
uncertain	3	1	5.0	10.0	70.0
disagree	4	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

SAME_MIS don't same mistakes

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	7	35.0	70.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

HARD_TRC hard to keep track which bits done

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	3	15.0	30.0	30.0
uncertain	3	4	20.0	40.0	70.0
disagree	4	2	10.0	20.0	90.0
strongly disagree	5	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

ANY_ORD problems any order

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	5.0	10.0	10.0
uncertain	3	4	20.0	40.0	50.0
disagree	4	5	25.0	50.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

DIFF_ROU unable different route

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	4	20.0	40.0	40.0
uncertain	3	5	25.0	50.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

SUGG_ROU prefer suggest. route

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	7	35.0	70.0	70.0
disagree	4	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

PROM_CLR prompts unclear from screen

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	5.0	10.0	10.0
disagree	4	4	20.0	40.0	50.0
strongly disagree	5	5	25.0	50.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

RET_LEFT prefer return I left off

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	7	35.0	70.0	70.0
agree	2	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

ON_TUTOR tutorials unclear from screen

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	20.0	40.0	40.0
agree	2	3	15.0	30.0	70.0
uncertain	3	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

LITT_UND little understanding of what doing

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	2	10.0	20.0	20.0
disagree	4	5	25.0	50.0	70.0
strongly disagree	5	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

OWN_PACE my own pace and direction

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	3	15.0	30.0	30.0
disagree	4	3	15.0	30.0	60.0
strongly disagree	5	4	20.0	40.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

PEN_PAPR prefer pencil & paper

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	3	15.0	30.0	50.0
uncertain	3	2	10.0	20.0	70.0
disagree	4	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

MRE_CONT prefer more control of time

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	20.0	40.0	40.0
agree	2	3	15.0	30.0	70.0
uncertain	3	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

TRY_DIFF learn try different things with system

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	3	15.0	30.0	30.0
disagree	4	4	20.0	40.0	70.0
strongly disagree	5	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

WILL_RET I will retain most of what I learnt

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	6	30.0	60.0	60.0
agree	2	3	15.0	30.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

ANS_PROB learnt lots problems

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	15.0	30.0	30.0
agree	2	6	30.0	60.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

PROB_QUK solve problem quick use this again

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	5	25.0	50.0	50.0
agree	2	5	25.0	50.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

MORE_HLP prefer more help

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	5.0	10.0	10.0
uncertain	3	7	35.0	70.0	80.0
disagree	4	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

FEW_HINT prefer few hints

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	3	15.0	30.0	50.0
uncertain	3	5	25.0	50.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

KEY_CORR which keys corresponded

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	1	5.0	10.0	10.0
disagree	4	5	25.0	50.0	60.0
strongly disagree	5	4	20.0	40.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

CLER_OPT clear which options allow select

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	7	35.0	70.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

GOT_LOST got lost because of options keys

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	1	5.0	10.0	10.0
disagree	4	6	30.0	60.0	70.0
strongly disagree	5	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

DIFF_COM different command - same key!

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
disagree	4	4	20.0	40.0	40.0
strongly disagree	5	6	30.0	60.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

EASYEXIT It was easy to exit

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	5	25.0	50.0	50.0
agree	2	4	20.0	40.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

NOT_SKIP couldn't skip some parts of exercise

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	4	20.0	40.0	40.0
uncertain	3	2	10.0	20.0	60.0
disagree	4	4	20.0	40.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

SIM_DELT simple to delete any answer

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	20.0	40.0	40.0
agree	2	6	30.0	60.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

RES_DATA not restart same data

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	4	20.0	40.0	40.0
agree	2	2	10.0	20.0	60.0
uncertain	3	2	10.0	20.0	80.0
disagree	4	1	5.0	10.0	90.0
strongly disagree	5	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

EASYCALC calculator easy to use

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	4	20.0	40.0	40.0
uncertain	3	2	10.0	20.0	60.0
disagree	4	4	20.0	40.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

HARDCALC complex calculator - hard to use

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	3	15.0	30.0	30.0
disagree	4	7	35.0	70.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

NOTCALC calculator does not do what you want

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	5.0	10.0	10.0
uncertain	3	4	20.0	40.0	50.0
disagree	4	4	20.0	40.0	90.0
strongly disagree	5	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

GRAPHUSE graph very useful

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	10.0	20.0	20.0
uncertain	3	6	30.0	60.0	80.0
disagree	4	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

GRAPHDIF difficult to read graph

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	4	20.0	40.0	40.0
disagree	4	6	30.0	60.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

GRAPHEAS easy compar. graph

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	5.0	10.0	10.0
uncertain	3	4	20.0	40.0	50.0
disagree	4	5	25.0	50.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

MORECOLR more graphics, colours & sounds

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	10.0	20.0	20.0
uncertain	3	7	35.0	70.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

EASY_COM easy communicate with computer

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	15.0	30.0	30.0
agree	2	6	30.0	60.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

LARG_INF confus. large info. on the screen

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	2	10.0	20.0	20.0
disagree	4	5	25.0	50.0	70.0
strongly disagree	5	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

EASY_SOL easy work out solution

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	15.0	30.0	30.0
agree	2	6	30.0	60.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

ENGH_INF not much info. on the screen

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	6	30.0	60.0	60.0
disagree	4	2	10.0	20.0	80.0
strongly disagree	5	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

ATTR_SCR attractive screen and clear

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	4	20.0	40.0	60.0
uncertain	3	4	20.0	40.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

CONF_SCR colour very confusing on screen

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	2	10.0	20.0	20.0
disagree	4	5	25.0	50.0	70.0
strongly disagree	5	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

FUN_USE comp. fun to use

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
agree	2	2	10.0	20.0	30.0
uncertain	3	2	10.0	20.0	50.0
disagree	4	3	15.0	30.0	80.0
strongly disagree	5	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

IMP_MORE prefer more improve.

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
agree	2	7	35.0	70.0	80.0
uncertain	3	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

ONE_BETT comp. much better than others

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	10.0	20.0	20.0
uncertain	3	5	25.0	50.0	70.0
disagree	4	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

SIM_READ simply read screen

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
uncertain	3	3	15.0	30.0	30.0
disagree	4	7	35.0	70.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

TOO_LONG prog. went too long

Value Label	Value	Frequency	Valid Percent	Cum Percent	Percent
strongly agree	1	1	5.0	10.0	10.0
agree	2	3	15.0	30.0	40.0
uncertain	3	2	10.0	20.0	60.0
disagree	4	3	15.0	30.0	90.0
strongly disagree	5	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

EASE_USE more use more easy

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	8	40.0	80.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

GIVE_ASS like assignments to be given

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	2	10.0	20.0	20.0
agree	2	5	25.0	50.0	70.0
uncertain	3	2	10.0	20.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

NOT_LERN didn't learn anything

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	3	15.0	30.0	30.0
agree	2	3	15.0	30.0	60.0
uncertain	3	3	15.0	30.0	90.0
disagree	4	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	

Valid cases 10 Missing cases 10

TROB_HLP helped me I trouble

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
agree	2	6	30.0	60.0	70.0
uncertain	3	1	5.0	10.0	80.0
disagree	4	2	10.0	20.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

GOOD_REV way to revise something forgotten

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	5	25.0	50.0	50.0
agree	2	4	20.0	40.0	90.0
uncertain	3	1	5.0	10.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

NOT_INTE student not use system

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	1	5.0	10.0	10.0
uncertain	3	6	30.0	60.0	70.0
disagree	4	3	15.0	30.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

WHY_CORR unclear why correct

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
agree	2	2	10.0	20.0	20.0
uncertain	3	3	15.0	30.0	50.0
disagree	4	5	25.0	50.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

MORE_EXA prefer more example

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
strongly agree	1	1	5.0	10.0	10.0
agree	2	9	45.0	90.0	100.0
	-1	10	50.0	Missing	
		-----	-----	-----	
	Total	20	100.0	100.0	
Valid cases	10	Missing cases	10		

APPENDIX R

APT Summaries of Responses to open Questions

What features did you find contributed to the STRENGTH of this system?

RESPONSE	FREQUENCY
Structured program, good teaching path, step by step approach, guided teaching.	4
Quick feedback, fast response.	2
Good instructions, helpful questions.	2
Easy to use.	1
Clear screen design.	1

What features did you find contributed to the WEAKNESS of this system?

RESPONSE	FREQUENCY
Lack of control, not interactive enough, inflexible learning.	4
Unfriendly error messages, unclear why wrong.	2
Weak help facility, lack of informative guidance.	2
Not interactive enough.	1
Too few examples.	1

If there was ONE outstanding improvement you could recommend to this system, what would it be?

RESPONSE	FREQUENCY
Better error messages, helpful messages, provide better feedback.	3
More detailed instructions, improved directions, better help.	3
Increase tutorial information.	2
Give further examples.	1
Make it more interactive.	1

APPENDIX S

APT Inter-Rater Reliability of Scores

- - Correlation Coefficients - -

NPRETSTK NPRETSTA

NPRETSTK	1.0000	1.0000
	(20)	(20)
	P= .	P= .000

NPRETSTA	1.0000	1.0000
	(20)	(20)
	P= .000	P= .

(Coefficient / (Cases) / 1-tailed Significance)

" . " is printed if a coefficient cannot be computed

- - Correlation Coefficients - -

POSTESTK POSTESTA

POSTESTK	1.0000	1.0000
	(20)	(20)
	P= .	P= .000

POSTESTA	1.0000	1.0000
	(20)	(20)
	P= .000	P= .

(Coefficient / (Cases) / 1-tailed Significance)

" . " is printed if a coefficient cannot be computed

APPENDIX T

APT Comparison of experimental and control differences using Independent T-tests

t-tests for Independent Samples of GROUP expt/contrl

Variable	Number of Cases	Mean	SD	SE of Mean

IMPROVEM				
GROUP >= 3	10	1.6750	2.741	.867
GROUP < 3	10	3.7000	1.522	.481

Mean Difference = -2.0250

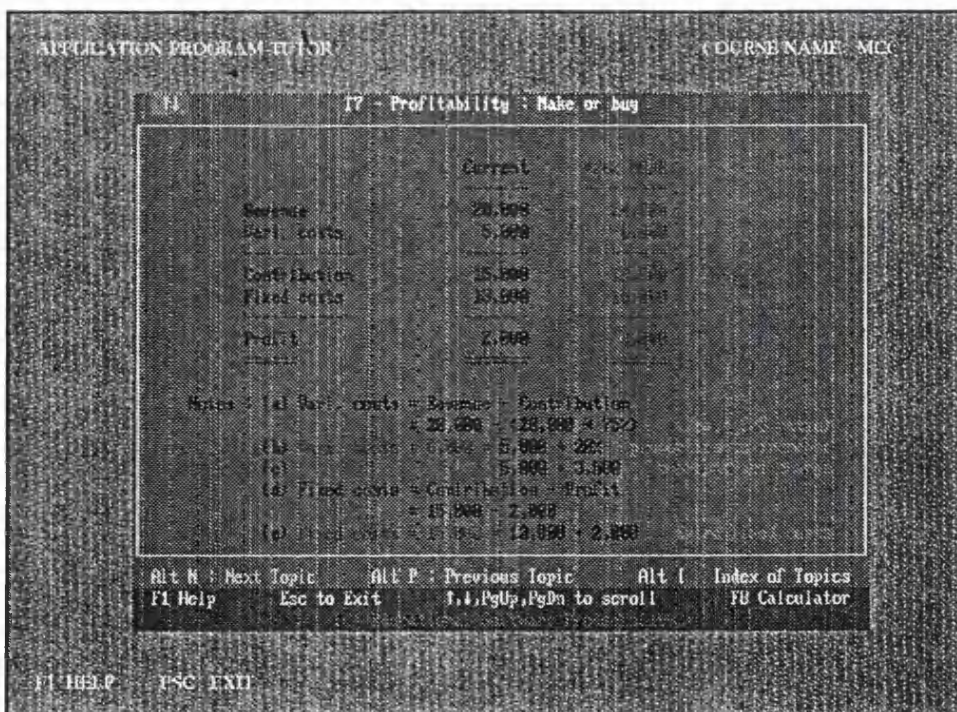
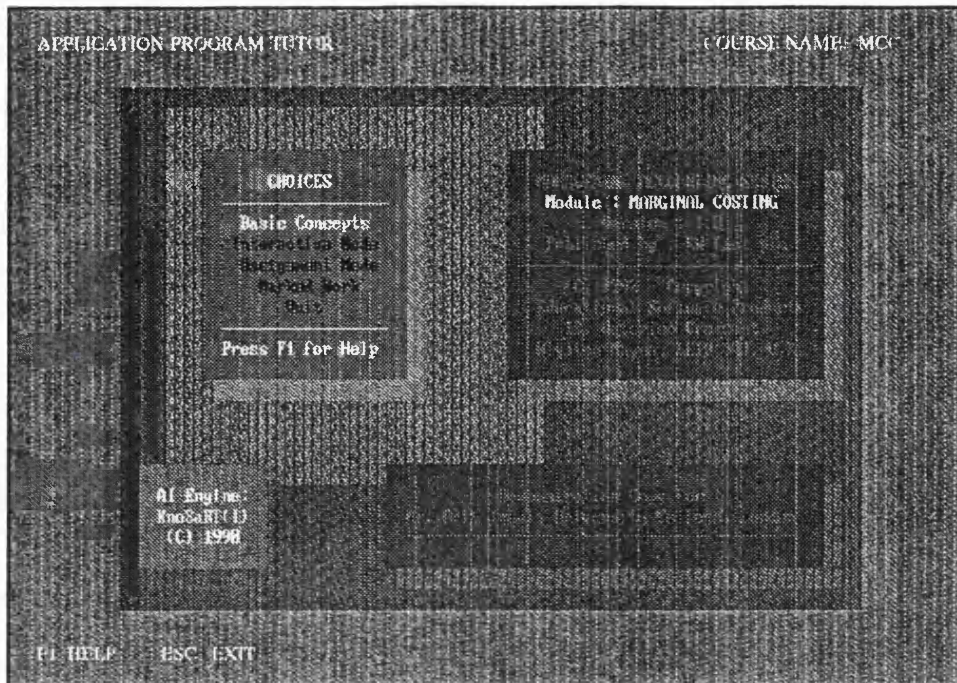
Levene's Test for Equality of Variances: F= .364 P= .554

t-test for Equality of Means					95%
Variances	t-value	df	2-Tail Sig	SE of Diff	CI for Diff

Equal	-2.04	18	.056	.992	(-4.108, .058)
Unequal	-2.04	14.07	.060	.992	(-4.151, .101)

APPENDIX U

APT Screen Dumps



APPLICATION PROGRAM TUTOR

COURSE NAME: MC

11

15 Profitability & Breaking even

The important things to note about this example are:

- i) The fixed costs do not alter when sales change
- ii) The break even point is when total revenue = total costs

Remember that this example was for a month's trading. The fixed costs are for given time period. Next month another set of fixed costs are incurred and another set of sales (and contribution) is required to pay for them. Strictly, the firm should allow breaks even at sales of 300 items a month.

The break even point can be identified in two ways, (1) by:

- a) At the sales of 300 needed to be sold to break even
- b) At the sales of 300 needed to break even

Alt N : Next Topic Alt P : Previous Topic Alt I : Index of Topics

F1 Help Esc to Exit F4, PgUp, PgDn to scroll F9 Calculator

F1 HELP ESC EXIT

APPLICATION PROGRAM TUTOR

COURSE NAME: MC

Computer generated example

Quantity 30000	Sale Price 7	Revenue	Cost - Variable 59.0000
	VC per unit	Total VC	
	Cost / Unit 59.0000	Total Cost	VC Ratio 2 59.00
		Fixed Cost	B/E Units 6125.66
		Profit	B/E Sales 42875.00
			B/S of Sales
			B/S of Units

Instructions

Please use F in I to move to the required line and enter a value in it.

Press Esc to exit.

Interactive messages

F1 Help F2 Erase entries F5 What-if F6 Graph F9 Calculator F10 Save

(C) Educational Software Products (1992)

F1 HELP ESC EXIT

APPLICATION PROGRAM TUTOR
COURSE NAME: MC

Computer generated example

Quantity	Sale Price	Revenue	Cost	Output
38887	100.0000	3888700		
VC per unit		Total VC		
41.0000		1591887		
Contribution		Total Cost	C/S Ratio	
59.0000		2289613	59.00	
		Fixed Cost	A/E Ratio	A/E Sales
		361414	6125.66	612566
		Profit	M/S Units	M/S Sales
		1928199	32681.34	3268134

Instructions

Please use 1 or 4 to move to the required box and enter a value in it.
Press Esc to quit

Interactive messages

Sorry! That isn't the correct answer.
Please try again! Suggested relationship:
M/S Sales = Revenue - E/E Sales

F1 Help F3 Erase entries F5 What-if F6 Graph F8 Calculator F9 Save
(C) Educational Software Products(1992)

F1 HELP ESC EXIT

APPLICATION PROGRAM TUTOR
COURSE NAME: MC

Computer generated example

Quantity	Sale Price	Revenue	Cost	Output
38887	100.0000	3888700		0.0000
VC per unit		Total VC		
41.0000		1591887		
Contribution		Total Cost	C/S Ratio	
59.0000		2289613	59.00	
		Fixed Cost	E/E Units	A/E Sales
		361414	6125.66	612566
		Profit	M/S Units	M/S Sales
		1928199	32681.34	3268134

Calculator

100
Calculator
1 2 3 4
5 6 7 8
9 0 + - * /
Esc to quit

Interaction messages

Pressing + or Enter again will transfer the result to Sale Price

F1 Help F3 Erase entries F5 What-if F6 Graph F8 Calculator F9 Save
(C) Educational Software Products(1992)

F1 HELP ESC EXIT

Main option : Interactive Mode

There are many interesting features in the software to improve your productivity and enhance the learning, including a pop-up calculator (you can see it right now if you press the F8 key).

FE Calculator

F4 HELP F5 ESC F6 EXIT

computer generated example

Computer generated example			
Quantity	Sales Price	Revenue	Cost
28257	98.0000	2769186	
MC per unit		Fixed MC	
7.000000		1381517	
Cost / Unit		Total Cost	MC Ratio
		Fixed Cost	MC Units
			MC & Sales
		Profit	MC Units
		1245418	MC & Sales
MC & Break Even			
MC & Break Even			

Instructions

Please use 1 or 1 in row
in the required box and
enter a value in it.
Press $\rightarrow$ for 10 or 100.

Interactive Messaging

You can calculate this free data on the screen.
Please try again!

F1 Help F3 Erase entries F5 What-if F6 Graph F8 Calculator tF8 Save
(C) Educational Software Products(1992)

F1 HELP TSC EXT)

14 - Profitability of Extra orders

This is particularly so if the firm has no spare capacity available. Spare capacity means that the firm has the ability to make more product without any increase in fixed costs. For example, the firm may have idle machines, or be able to run an extra production shift.

The question for the firm is whether (from the profit point of view - there are others) it is worth taking such an order at a lower than normal selling price?

Example 1

normal price per unit of product = £4,
normal output in 2000 = 1000 units per month,
total variable cost per unit = £3.

Measurements plc. offers us a working order of 1000 units at a reduced price of £2 per unit. Do we have the necessary spare capacity to take the order worth taking?

Hit M : Next Topic Hit P : Previous Topic Hit I : Index of Topics
F1 Help Esc to Exit F1,F9,PgUp,PgDn to scroll F8 Calculator

F1 HELP ESC EXIT

16 - Profits and Margin of safety

To calculate the margin of safety (M/S) of a business, all you have to do is to deduct the B/S revenue from its target revenue - calculate the sales required to generate a contribution equal to the target profit.

As a formula :

$$M/S \text{ Sales revenue} = \text{Total Sales revenue} - B/S \text{ Sales revenue}$$

or

$$M/S \text{ Sales revenue} = \text{Profit} / C/S \text{ ratio}$$

In the above example, we had

	Ca. A	Ca. B
Revenue	24,000	24,000
Variable costs	(5,000)	(4,500)
Contribution margin	19,000	19,500
C/S ratio	79.17%	81.25%
Fixed costs	(4,000)	(4,000)
Target contribution	4,000	4,000

Hit M : Next Topic Hit P : Previous Topic Hit I : Index of Topics
F1 Help Esc to Exit F1,F9,PgUp,PgDn to scroll F8 Calculator

F1 HELP ESC EXIT

REFERENCES

Adelman, Leonard (1992) "Evaluating Decision Support and Expert Systems", New York: John Wiley and Sons, Inc.

Altman, D.G. (1991) "Practical Statistics for Medical Research" 1st ed., London: Chapman and Hall, ISBN 0412276305.

Anderson, John R., Boyle, C. Franklin and Reiser, Brian J. (1985) Intelligent Tutoring Systems, "Science", "228", p456-462.

Angell, Ian O. and Smithson, Steve (1991) "Information Systems Management: Opportunities and Risks", Basingstoke: The Macmillian Press Ltd.

Atkinson, G.B.J. (1983) "The Economics of Education", London: Hodder and Stoughton.

Baker, E. L. (1991) in Burns, H., Parlett, J. W. and Redfield, C. L. (eds.) "Intelligent Tutoring Systems - Evolutions in Design", Lawrence Erlbaum Associates, New Jersey.

Barker, John and Tucker, Richard N. (Eds) (1990) "The Interactive Learning Revolution: Multimedia in Education and Training", London: Kogan Page Ltd.

Basili, Victor R. (1995) Applying the Goal/Question/Metric Paradigm in the Experience Factory. In Fenton, Norman., Whitty., Robin and Iizuka, Yoshinori (Eds.) "Software Quality Assurance and Measurement A Worldwide Perspective", London: International Thomson Computer Press.

Bawden, David (1990) "User-oriented Evaluation of Information Systems and Services", Aldershot: Gower Publishing Company Limited.

Bell, Judith (1987) "Doing Your Research Project", Stony Stratford: Open University Press.

Bell, Margaret (1985) The Coventry Computer Based Learning Project 1982-1983, In Duncan, K.A. and Harris, D.I. (eds) "Computers in Education", Amsterdam: Elsevier Science publishers.

Bentley, C. (1982) "Computer Project Management". London: Heyden Son Ltd.

Bento, Simon (1990) A Holistic Model for the Evaluation of Intelligent Tutoring Systems, "Interactive Learning International", "6", p173-180.

Brooks, F.P. (1975) "The Mythical Man-Month", Reading: Addison-Wesley.

Brown, Geoffrey (1989) "Minds, Brains and Machines", Bristol: Bristol Classical Press.

Bryman, Alan and Cramer, Duncan (1994) "Quantitative Data Analysis for Social Scientists" London: Routledge.

Campbell, J.A. (1986) On Artificial Intelligence, "Artificial Intelligence Review", "1", p3-9.

Carbonell, Jamie R. (1970) "Mixed-Initiative Man-Computer Instructional Dialogues". PhD Dissertation. Massachusetts Institute of Technology, Cambridge, M.A.

Castellan, N. John (1993) Evaluating Information Technology in Teaching and Learning, "Behavior Research Methods, Instruments and Computers", "25", No. 2, p233-237.

Chuen, Lee Moon (1989) "ICAI Authoring System Development: A Multi-Paradigm Approach", PhD Thesis, University of London.

Clancey, William J. (1987) "Knowledge-Based Tutoring - The GUIDON Program", Cambridge, Massachusetts: The MIT Press.

Cleland, D.I. (1990) "Project Management - Strategic Design and Implementation". Blue Ridge Summit: Tab Books Inc.

Clifton, H.D. (1986) "Business Data Systems", New Jersey: Prentice-Hall International.

Cohen, Paul R. and Howe, Adele E. (1989) Toward AI Research Methodology: Three Case Studies in Evaluation, "IEEE Transactions on Systems, Man, and Cybernetics", "19", No. 3, p634-646.

Collins, Harry (1992) Will Machines Ever Think?, "New Scientist", "June", p36-40.

Cook, John (1991) Evaluation of an Intelligent Tutoring System, "MSc Thesis", City University, England, September 1991.

Cox, Brad J. (1995) There is a Silver Bullet. In Heap, Nick., Thomas, Ray., Einon, Geoff., Mason, Robin and Mackay, Hughie (Eds.) "Information Technology and Society", London: Sage Publications Ltd.

Cox, D.R. (1958) "Planning of Experiments". London: John Wiley & Sons, Inc.

Davis, Fred D. and Venkatesh, Viswanath (1996) A Critical Assessment of Potential Measurement Biases in the Technology Acceptance Model: Three Experiments, "International Journal of Human-Computer Studies", "45", p19-45.

Dell, Michael's article is in "Fortune" pub. October 7, 1991.

"Dictionary of Quotations" (1990), London: Bloomsbury Publishing Limited. ISBN-1-85605-080-7. (12;5;6;9;5;12;4)

Downton, Andy (Ed.) (1993) "Engineering The Human-Computer Interface", London: McGraw-Hill Book Company.

Duncan, Nancy C. (1993) Evaluation of Instructional Software: Design Considerations and Recommendations, "Behavior Research Methods, Instruments and Computers", "25", No. 2, p223-227.

Dungworth, Steve and Russell, David (1994) The evolution of interactive CAL in management accounting: Further evaluation of a management accounting in practice, "5th Annual CTI-AFM Conference", April 94, Coventry.

Ebel, Robert L. and Frisbie, David A. (1991) "Essentials of Educational Measurement", New Jersey: Prentice-Hall.

Edwards, John S. (1991) "Building Knowledge-Based Systems: Towards a Methodology", London: Pitman Publishing.

Elmes, David G., Kantowitz, Barry H. and Roediger, Henry L. (1985) "Research Methods in Psychology", Minnesota: West Publishing Company.

Evans, C.D., Meek, B.L. and Walker, R.S. (1993) "User Needs in Information Technology Standards", Oxford: Butterworth-Heinemann Ltd.

"Every Child in Britain - Report of the Channel Four Commission on Education", published in 1991, by Channel 4 Television, London. ISBN 1-85144-018-6. ***

Faint, Richard (1994) "Technologies in Learning: A Guide to the effective application of learning technology", Glasgow: SCET (Employment Department).

Farr, A. (1984) "Establishing a Quality Control Circle Programme" in Sasaki, N. and Hutchins, D. (eds) "The Japanese Approach to Product Quality". Oxford: Pergamon Press.

Fenton, N.E. (1991) "Software Metrics - A Rigorous Approach", London: Chapman & Hall.

Ford, Lindsey (1988) in Self, John "Artificial Intelligence and Human Learning", London: Chapman and Hall.

French, C.S. (1984) "Computer Science: An Instructional Manual", Eastleigh: D.P. Publications Ltd.

Frith, D. S. and Macintosh, H. G. (1984) "A Teacher's Guide to Assessment", Cheltenham: Stanley Thornes (Publishers) Ltd.

Gillies, Alan C. (1992) "Software Quality: Theory and Management", London: Chapman and Hall.

Goodyear, Peter (1991) in Goodyear, P. (ed.) "Teaching Knowledge and Intelligent Tutoring", New Jersey: Ablex Publishing Corporation.

Greene, Judith and D'Oliveira, Manuela (1993) "Learning to Use Statistical Tests in Psychology", Buckingham: Open University Press.

Guthrie, Bart Matthew and McPherson, Moira (1992) The Efficacy of a Customized Approach to Computer Assisted Instruction, "Journal of Computer-Based Instruction", "19", No. 3, p100-104.

Hannafin, Michael J. and Peck, Kyle L. (1988) "The Design, Development, and Evaluation of Instructional Software", New York: Macmillan Publishing Company.

Hawkrigde, David (1992) "Feel the Width: Quality in Modern Technology", The Open University, Milton Keynes: Centre for Information Technology in Education Report No. 159.

Hawkrigde, David (1991) Challenging Educational Technology, "Educational and Training Technology International", "28", No. 2, p102-110.

Hays, Patrick J., Ford, Kenneth M. and Adams-Webber, Jack R. (1992) Human Reasoning about Artificial Intelligence, "Journal of Experimental and Theoretical Artificial Intelligence", "4", p247-263.

Hays, William L. (1981) "Statistics" New York: Holt, Rinehart and Winston, Inc.

Herman, Joan L., Morris, Lynn Lyons and Fitz-Gibbon, Carol Taylor (1987) "Evaluator's Handbook", Newbury Park: Sage Publications, Inc.

Hooper, Kristina (1990) HyperCard: A Key to Educational Computing. In Ambron, Sueann and Hooper, Kristina (1990) "Learning with Interactive Multimedia: Developing and Using Multimedia Tools in Education", Washington: Microsoft Press.

Hopkins, David (1989) "Evaluation for School Development", Milton Keynes: Open University Press.

Howard, Stephen and Murray, Dianne M. (1987) "A Survey and Classification of Evaluation Techniques for HCI" NPL Technical Report DITC 94/87 National Physical Laboratory: Teddington, Middlesex.

Hutchins, D. (1984) "Quality Circles - The Missing Link" in Sasaki, N. and Hutchins, D. (eds) "The Japanese Approach to Product Quality". Oxford: Pergamon Press.

Iqbal, Arif (1990) Computer Technology - Is it a Solution to the American Education Crisis? Internal Note, "IN/92/4", Department of Computer Science, University College London, 6th May 1990.

Iqbal, Arif., Oppermann, Reinhard., Patel, Ashok and Kinshuk (1999) The Role of Quality Metrics in the Development of Intelligent Tutoring Systems, "CAL 99 - Virtuality in Education", 28-31st March, 1999, London, UK.

Iqbal, Arif., Oppermann, Reinhard., Patel, Ashok and Kinshuk (1999) A Classification of Evaluation Methods for Intelligent Tutoring Systems, "Software-Ergonomie '99", 8-11th March, 1999, Walldorf/Baden, Germany.

Iqbal, Arif., Patel, Ashok and Kinshuk (1998) "Use of Quality Metrics in User-Oriented Evaluation Standards Development for Computer Assisted Instruction Systems". Computers in Psychology Conference, 15-17th April, 1998, University of York, England, U.K.

Jackson, P.W. (1966) "The teacher and the machine: Observations on the impact of educational technology", unpublished manuscript, University of Chicago. In Morrison, A. and McIntyre, D. (1972) "The Social Psychology of Teaching", Harmondsworth: Penguin Books Ltd.

Jones, Barry T., and McCormac, Kathleen (1992) Empirical evidence shows that measuring users' opinions is not a satisfactory way of evaluating computer-assisted learning in nurse education, "international Journal of Nursing Studies, Vol 29 (4), p411-425.

Jordan, Eleanor W. and Machesky, Jefry J. (1990) "Systems Development: Requirements, Evaluation, Design, and Implementation", Boston: PWS-Kent Publishing Co.

Juristo, Natalia., Mate, Luis and Tovar, Edmundo (1997) A Measurement Framework that supports Evaluation of Knowledge Engineering Products, "Experts Systems", "14", 3, p140-153.

Kaye, Anthony (1995) Computer Supported Collaborative Learning. In Heap, Nick., Thomas, Ray., Einon, Geoff., Mason, Robin and Mackay, Hughie (Eds) "Information Technology and Society", London: Sage Publications Ltd.

Kemp, D.A. (1988) "Computer-Based Knowledge Retrieval", London: The Association for Information Management.

Kerlinger, Fred N. (1973) "Foundations of Behavioral Research", New York: Holt, Rinehart and Winston, Inc.

Kerr, Stephen T. (1989) Pale Screens: Teachers and Electronic Texts. In Jackson, Philip W. and Haroutunian-Gordon, Sophie (Eds.) "From Socrates to Software: The Teacher as Text and The Text as Teacher", Chicago, Illinois: The National Society for the Study of Education.

Kidder, Louise H. (1981) "Selltiz, Wrightsman and Cook's Research Methods in Social Relations", New York: Holt, Rinehart and Winston, Inc.

Kimball, Ralph (1982) in Sleeman, D. and Brown, J.S. (eds) "Intelligent Tutoring Systems", London: Academic Press Ltd.

Kinnear, Paul R. and Gray, Colin D. (1994) "SPSS for Windows Made Simple", Hove: Lawrence Erlbaum Associates, Publishers.

Knapper, Christopher Kay (1980) "Evaluating Instructional Technology", Halstead Press, New York.

Kinshuk (1996) "Computer Aided Learning For Entry Level Accountancy Students", PhD Thesis, De Montfort University, England, July 1996.

Koestler, Arthur (1989) "The Ghost in the Machine", London: Arkana.

Knirk, Frederick G. and Gustafson, Kent L. (1986) "Instructional Technology: A Systematic Approach to Education" New York: Holt, Rinehart and Winston.

Kulik, James A. (1994) Meta-Analytic Studies of Findings on Computer-Based Instruction. In Baker, Eva L. and O'Neil, Harold F. (Eds.) "Technology Assessment in Education and Training", Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Kumar, K. (1990) Post Implementation Evaluation of Computer-Based Information Systems: Current Practices. "Communications of the ACM", "33", p203-212.

Legree, Peter J., Gillis, Philip D. and Orey, Michael A. (1993) The Quantitative Evaluation of Intelligent Tutoring System Applications: Product and Process Criteria, "Journal of Artificial Intelligence and Education", "4", 2/3, p209-226.

Legree, Peter J. and Gillis, Philip D. (1991) Product Effectiveness Evaluation Criteria for Intelligent Tutoring Systems, "Journal of Computer-Based Instruction", "18", 2, p57-62.

Lindgaard, Gitte (1994) "Usability Testing and System Evaluation: A Guide for Designing Useful Computer Systems", London: Chapman & Hall.

Littman, D. and Soloway, E. (1988) in Polson, M.C. and Richardson, J.J. (eds.) "Foundations of Intelligent Tutoring Systems", New Jersey: Lawrence Erlbaum Associates Publishers.

Lo, Leo Man Pong (1991) "The Evolution of Computer Assisted Learning from Instructional Tool to Cognitive Tool: Perspective from the Psychology of Learning with Special Reference to Science Education". M.A. thesis, Institute of Education, University of London, 1991.

Lock, D. (1988) "Project Management", Aldershot: Gower Pub Co. Ltd.

Lucey, T. (1992) "Management Accountancy", London:D.P. Publications Ltd.

Mark, Mary A. and Greer, Jim E. (1993) Evaluation Methodologies for Intelligent Tutoring Systems, "Journal of Artificial Intelligence and Education", "4", 2/3, p129-153.

Martin, Paul and Bateson, Patrick (1986) "Measuring Behaviour: An Introductory Guide" Cambridge: Cambridge University Press.

McGraw, Karen L. (1992) "Designing and Evaluating User Interfaces for Knowledge-Based Systems", New York: Ellis Horwood Ltd.

Moller, K.H. and Paulish, D.J. (1993) "Software Metrics", London: Chapman and Hall.

Morris, Michael G. and Dillon, Andrew P. (1996) The Importance of Usability in the Establishment of Organizational Software Standards for End User Computing, "International Journal of Human-Computer Studies", "45", p243-258.

Mott, Graham (1984) "Accountancy for Non-Accountants", London: Pan Books Ltd.

Mousinho, G. (1990) "Runaway! (Project Management)". "Systems International", 18, 6, p35-40.

Morris, Nigel (1990) "Understanding Educational Evaluation", London: Kogan Page Ltd.

Murray, Tom (1993) Formative Qualitative Evaluation for 'Exploratory' ITS Research, "Journal of Artificial Intelligence and Education", "4", 2/3, p179-207.

Marshall, David (1988) "CAL/CBT - The Great Debate", Bromley, Kent: Chartwell-Brett (Publishing and Training) Ltd.

Neisser, Ulric (1963) The Imitation of Man by Machine, "Science", "139", p193-197.

Norris, Mark., Rigby, Peter and Payne, Malcolm (1993) "The Healthy Software Project: A Guide to Successful Development and Management", Chichester: John Wiley and Sons.

Nwana, Hyacinth Sama (1993) in Nwana H.S. (ed.) "Mathematical Intelligent Learning Environments", Oxford: Intellect Books.

Nwana, H.S. (1990) Intelligent Tutoring Systems: An Overview, "Artificial Intelligence Review", "4", p251-277.

O'Keefe, R. M. and O'Leary, D. (1993) Expert System Verification and Validation: a Survey and Tutorial, "Artificial Intelligence Review", "7", p3-42.

Oppenheim, A.N. (1986) "Questionnaire Design and Attitude Measurement", Aldershot: Gower Publishing Company Ltd.

O'Shea, Tim and Self, John (1983) "Learning and Teaching with Computers", Brighton: The Harvester Press.

Or-Bach, Rachel and Bar-On, Ehud (1993) 'TALK'ing About Evaluation, "Journal of Artificial Intelligence and Education", "4", 2/3, p227-243.

Parlett, M. and Hamilton, D. (1972) Evaluation as Illumination: A new Approach to the Study of Innovative Programmes. Occasional Paper No. 9, Centre for Research in the Educational Sciences, University of Edinburgh. Reprinted in Hamilton, D. et al. (eds) (1977) "Beyond the Numbers Game", London: Macmillan.

Patel, A. and Kinshuk (1996) Intelligent Tutoring Tools: Redesigning ITSs for Adequate Knowledge Transfer Emphasis, "International Conference on Intelligent and Cognitive Systems - ICICS '96", Tehran, Iran, 23-23 September 1996. In Press.

Perciful, E.G. (1993) Curriculum planning and computer-assisted instruction (CAI) within clinical nursing education, "Sixteenth Annual Symposium on Computer Applications in Medical Care", p414-18, Baltimore, MD., USA, 8-11 Nov. 1992.

Pirsig, Robert M. (1974) "Zen and the Art of Motorcycle Maintenance: An Inquiry into Values", London: Vintage.

Preston, J.M. (1992) "Compact Disc-Interactive: A Designer's Overview", Deventer, The Netherlands: Kluwer Technical Books.

Ransdell, Sarah (1993) Educational Software Evaluation Research: Balancing Internal, External, and Ecological Validity, "Behavior Research Methods, Instruments and Computers", "25", No. 2, p228-232.

Reiser, Robert A. and Kegelmann, Harald W. (1994) Evaluating Instructional Software: A Review and Critique of Current Methods, "Educational Technology Research and Development", "42", 3, p63-9.

Report of the Paul Hamlyn Foundation, National Commission on Education (1993) "Learning to Succeed", London: William Heinemann Ltd. ***

Reynolds, Chris (1990) Moving Information Technology Research from the Laboratory to the Classroom, In Cott, Diana East., Farmer, Bob and Lantz, Brian "Aspects of Educational and Training Technology, Volume XXIII: Making Learning Systems Work", London: Kogan Page Ltd.

Rich, Tom (1991) Computer Technology and Education: Past Performance and Future Promise, "Educational and Training Technology International", "28", No. 2, p147-153.

Richer, Mark H. and Clancey, William J. (1987) GUIDON_WATCH: A Graphic Interface for Viewing a Knowledge-based system, in Lawler, Robert W. and Yazdani, Masoud (eds) "Artificial Intelligence and Education: Volume One", New Jersey: Ablex Publishing Corporation.

Robson, Colin (1983) "Experiment Design and Statistics in Psychology", Harmondsworth: Penguin Books Ltd.

Robson, M. (1982) "Quality Circles - A Practical Guide". Aldershot: Gower Pub. Co. Ltd.

Rogers, Glyn and Badham, Linda (1992) "Evaluation in Schools: Getting Started on Training and Implementation", London: Routledge.

Romiszowski, A.J. (1988) "The Selection and Use of Instructional Media: For improved classroom teaching and for interactive, individualized instruction", London: Kogan Page Ltd.

Romiszowski, A.J. (1981) "Designing Instructional Systems: Decision Making in Course Planning and Curriculum Design", London: Kogan Page Ltd.

Ross, Peter (1987) Intelligent Tutoring Systems, "Journal of Computer-Assisted Learning", "3", 4, p194-203.

Roszak, Theodore (1986) "The Cult of Information: The Folklore of Computers and the True Art of Thinking", London: Paladin Books.

Rust, John and Golombok, Susan (1989) "Modern Psychometrics: The Science of Psychological Assessment", London: Routledge.

Schank, Roger (1991) Where's the AI?, "AI Magazine", "12", No. 4, p38-49.

Schumacher, E.F. (1973) "Small is Beautiful - A Study of Economics as if People Mattered", London: Sphere Books Ltd.

Scriven, M. (1991) Beyond Formative and Summative Evaluation. In M.W. McLaughlin and D.C. Philips (Ed) "Evaluation and Education: At Quarter Century", p19-64, Chicago, IL: National Society for the Study of Education.

Sculley, John (1993) in Papert, Seymour "Mindstorms: Children, Computers, and Powerful Ideas", Hemel Hempstead: Harvester Wheatsheaf.

Seffah, Ahmed (1997) An Object-Oriented Framework for Modeling and Designing Intelligent Training Systems, "Software - Practice and Experience", Vol 27 (10), p1233-1243.

Self, J. (1985) A perspective on intelligent computer-assisted learning, "Journal of Computer Assisted Learning", "1", p159-166.

Self, J.A. (1974) Student models in computer-aided instruction, "International Journal of Man-Machine Studies", "6", p261-276.

Semper, Rob (1990) HyperCard and Education: Reflections on the HyperBoom. In Ambron, Sueann and Hooper, Kristina (1990) "Learning with Interactive Multimedia: Developing and Using Multimedia Tools in Education", Washington: Microsoft Press.

Shackel, Brian (1996) Ergonomics: Scope, Contribution and Future Possibilities. "The Psychologist" July, Vol 9, No. 7, p304-308.

Shea, Gregory P. (1991) Quality Circles: the Danger of Bottled Change. In Henry, Jane and Walker, David., "Managing Innovation", London: Sage Publications Ltd.

Shute, V.J. and Regian, J.W. (1993) Principles for Evaluating Intelligent Tutoring Systems, "Journal of Artificial Intelligence and Education", "4", 2/3, p245-271.

Snodgrass, Joan Gay (1977) "The Numbers Game: Statistics for Psychology", Oxford University Press, Inc.

Snow, C.P. (1964) "The Two Cultures", Cambridge: Cambridge University Press.

Snyder, Tom and Palmer, Jane (1986) "In Search of The Most Amazing Thing: Children, Education and Computers", Reading: Massachusetts: Addison-Wesley Pub. Co., Inc.

Sommerville, Ian (1989) "Software Engineering", Wokingham: Addison-Wesley Pub. Co.

Sperling, Abraham (1982) "Psychology", London: William Heinemann Ltd.

Stalley, B. (1992) A wider view can show the value of software, "Computer Weekly", August 27.

Stebbing, L.(1989) "Quality Assurance - The Route to Efficiency and Competitiveness". Chichester: Ellis Horwood Ltd.

Stenhouse, L. (1975) "An Introduction to Curriculum Research and Development", London: Heinemann Educational.

Stones, Edgar (1983) "Psychology of Education: A Pedagogical Approach", London: Methuen and Co. Ltd.

Taylor, Josie and Laurillard, Diana (1995) Supported Resource Based Learning. In Heap, Nick., Thomas, Ray., Einon, Geoff., Mason, Robin and Mackay, Hughie (Eds) "Information Technology and Society", London: Sage Publications Ltd.

Thomas, David L. and Bostow, Darrel E. (1991) Evaluation of Pre-Therapy Computer-Interactive Instruction, "Journal of Computer-Based Instruction", "18", No. 2, p66-70.

Tonge, Alyson., Beacham, Nigel and Gallagher, Denise (1994) The Implementation and Evaluation of Computer Based Learning in Accounting Education, "The International Symposium on Independent Learning and Flexible Study", Cambridge, September 1994.

Tonge, Alyson., Johnston, Karen., Dungworth, Steve and Russell, David (1993) "The only truly effective CAL is interactive" - An evaluation of a management accounting package in practice, "4th Annual CTI-CFM Conference", March 93, Leicester.

Twidale, Michael (1993) Redressing the Balance: The Advantages of Informal Evaluation Techniques for Intelligent Learning Environments, "Journal of Artificial Intelligence and Education", "4", 2/3, p155-178.

Tyler, Leona E. (1973) "Tests and Measurements", London: Prentice-Hall International, Inc.

Underwood, Jean D. and Underwood, Geoffrey (1990) "Computers and Learning: Helping children acquire thinking skills", Oxford: Basil Blackwell Ltd.

Yazdani, Masoud (1987) Intelligent Tutoring Systems: An Overview, in Lawler, Robert W. and Yazdani, Masoud (eds) "Artificial Intelligence and Education: Volume One", New Jersey: Ablex Publishing Corporation.

Yazdani, M. (1986) Intelligent tutoring systems survey, "Artificial Intelligence Review", "1", 1, p43-52.

Watson, D.M. (1993) Classroom processes and geography simulations, "Computers and Education", July-Sept, Vol. 21(2), p9-16.

Welsh, Josephine A. (1993) The Effectiveness of Computerized Instruction at the College Level: Five Suggestions for Successful Implementation, "Behavior Research Methods, Instruments, and Computers", "25", (2), p220-222.

Westrum, Ron (1991) "Technologies and Society: The Shaping of People and Things", Belmont, California: Wadsworth Publishing Company.

Wilkinson-Riddle, G.J. and Patel, A. (1993) Human Tutor Emulation in Teaching Numerical Business Subjects - a Software Breakthrough, "Proceedings of the tenth International Conference on Technology and Evaluation, March 21-24, p174-176.

Wilson, John R. and Corlett, E. Nigel (Edited) (1990) "Evaluation of Human Work - A Practical Ergonomics Methodology" London: Taylor & Francis Ltd.

Winne, Philip H. (1993) A Landscape of Issues in Evaluating Adaptive Learning Systems, "Journal of Artificial Intelligence and Education", "4", 4, p309-332.

Wright, George and Fowler, Chris (1986) "Investigative Design and Statistics", Harmondsworth, Middlesex: Penguin Books Ltd.

Wyer, Jo-Anne (1987) in Rushby, Nick (eds.) "Technology Based Learning: Selected Readings", London: Kogan Page Ltd.

Worthen, Blaine R. and Sanders, James R. (1987) "Educational Evaluation - Alternative Approaches and Practical Guidelines", New York: Longman Inc.

Yeates, D. (1991) "Project Management for Information Systems". London: Pitman Publishing.

Yourdon, Edward (1996) When Good Enough is Best, "Byte", "21", 9, p85.

BIBLIOGRAPHY

Barber, Theodore Xenophan (1976) "Pitfalls in Human Research: Ten Pivotal Points". New York: Pergamon Press, Inc.

Buckle, J.K. (1977) "Managing Software Projects". Malabar, Florida: Robert E. Krieger Pub. Co., Inc.

Cleland, D.I. and King, W.R. (1986) "Systems Analysis and Project Management". New York: McGraw-Hill Book Co.

Collins, T. (1992) "Sunrise Industry" in "Computer Weekly" Pub. January 23, 1992.

Gildersleeve, T.R. (1985) "Data Processing Project Management". New York: Van Nostrand Reinhold Co. Inc.

Rose, M. (1978) "Industrial Behaviour - Theoretical Development since Taylor". Harmondsworth: Penguin Books Ltd.

Tasker, P. (1987) "Inside Japan - Wealth, Work and Power in the New Japanese Empire". Harmondsworth: Penguin Books Ltd.

N.B. *** = indicates no author(s) were identified within the reference material.