

ON COMPUTER-BASED ASSESSMENT OF MATHEMATICS

DANIEL ARTHUR PEAD, BA

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

December 2010

Abstract

This work explores some issues arising from the widespread use of computer based assessment of Mathematics in primary and secondary education. In particular, it considers the potential of computer based assessment for testing “process skills” and “problem solving”. This is discussed through a case study of the *World Class Tests* project which set out to test problem solving skills.

The study also considers how on-screen “eAssessment” differs from conventional paper tests and how transferring established assessment tasks to the new media might change their difficulty, or even alter what they assess. One source of evidence is a detailed comparison of the paper and computer versions of a commercially published test – Nelson's *Progress in Maths* - including a new analysis of the publisher's own equating study.

The other major aspect of the work is a design research exercise which starts by analysing tasks from Mathematics GCSE papers and proceeds to design, implement and trial a computer-based system for delivering and marking similar styles of tasks. This produces a number of insights into the design challenges of computer-based assessment, and also raises some questions about the design assumptions behind the original paper tests. One unanticipated finding was that, unlike younger pupils, some GCSE candidates expressed doubts about the idea of a computer-based examination.

The study concludes that implementing a Mathematics test on a computer involves detailed decisions requiring expertise in both assessment and software design, particularly in the case of richer tasks targeting process skills. It concludes with the proposal that, in contrast to its advantages in literacy-based subjects, the computer may not provide a “natural medium for doing mathematics”, and instead places an additional demand on students. The solution might be to reform the curriculum to better reflect the role of computing in modern Mathematics.

Acknowledgements

I would like to offer my sincere thanks to the following, without which this work would have been impossible:

The teachers and pupils in the schools who voluntarily assisted in the trials which form a crucial part of this work.

Pat Bishop and Rosemary Bailey, who lent their professional experience to the marking of the GCSE-level trials.

Carol Craggs for support during the observations of *Progress in Maths*.

Sean McCusker for his help with an awkward Rasch.

QCA and nferNelson for funding projects which made parts of the work possible.

My colleagues at the University of Nottingham.

Table of Contents

Chapter 1: Outline of the study

1.1: Rationale.....	1
1.2: Research Questions.....	3
1.3: Outline of the work	5
Testing “higher order” or “problem solving” skills	5
Computerising existing tests	6
Analysing GCSE with a view to computer delivery	8
Computerising GCSE mathematics – a design research experiment	8
The author's contribution to this work.....	9

Chapter 2: Key issues in Mathematics Assessment

2.1: Introduction.....	11
2.2: The influence of tests on the taught curriculum.....	12
2.3: Achieving “balance” in assessment – a case study	13
2.4: Assessing Problem Solving, Functional Mathematics and Process Skills . . .	18
Problem solving vs. “word problems”	21
2.5: The challenges of computerisation.....	22
Is mathematics “harder” on computer?.....	22
Capturing working and “method marks”	23
A natural medium for “doing mathematics”?	23
Automatic marking.....	25
2.6: The implications of eAssessment for assessment reform	25
2.7: A note on “formative assessment”.....	27

Chapter 3: Assessing problem solving skills: a case study

3.1: Introduction.....	29
------------------------	----

Table of Contents

3.2: The World Class Tests project	30
The brief.....	30
Educational principles.....	30
3.3: The role of the computer.....	34
3.4: Illustrative examples of tasks.....	36
Simulated experiments and “microworlds”	36
Mathematical games.....	38
Exploring rich data sets.....	39
Use of the workbooks.....	41
3.5: The development process.....	46
Initial design.....	46
Specification, commissioning and implementation	46
Trial and refinement.....	48
Some design challenges.....	49
3.6: Technical and Logistical Challenges.....	50
Technical issues.....	50
Project management issues.....	51
3.7: Outcome of the project.....	52
3.8: Conclusions.....	53

Chapter 4: Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

4.1: Introduction.....	55
4.2: The Progress in Maths tests.....	56
4.3: Approach.....	58
Analysis of the equating study data.....	58
Design critique of the questions.....	59
School observations.....	60
Surveys and interviews.....	61
Relationship between the approaches.....	62
4.4: Results of the equating study for PIM6 and PIM7	62
Is there an ability effect?.....	62
Is there a school effect?.....	64
Is the order of administration causing an effect?	66
Was there a task effect?.....	68
Statistical tests of item differences.....	73
Use of advanced scaling techniques.....	79
Conclusions from the nferNelson Equating Study	81
4.5: Task design critique & school observations.....	83
Presentation of spoken prompts.....	83
Pictures and distraction.....	86
Colour and accessibility.....	87
Changes in predicted difficulty.....	88
Validity of rich contexts and problem solving.....	91
Some examples of design issues.....	95
4.6: Schools and equipment provision.....	102
4.7: Weaknesses of the experimental model.....	103

Table of Contents

4.8: Conclusions.....	104
Implications for our research questions.....	104
Expectations of equivalence.....	104
Design guidelines for equivalence.....	105
The importance of close observations.....	106
Chapter 5: Computerising mathematics assessment at GCSE: the challenge	
5.1: Introduction.....	107
5.2: The current state of GCSE mathematics.....	108
Scope of this work.....	108
A typical GCSE mathematics examination.....	108
Responses and marking.....	110
5.3: Weaknesses of the GCSE format.....	111
Fragmentation.....	111
Technical vs. strategic skills.....	113
Mathematics for an IT-driven world.....	114
Plausibility of “realistic” concepts.....	115
Mathematical validity.....	118
Allocation of marks.....	122
Heterogeneous testing.....	124
5.4: Adapting a GCSE-style paper task for computer.....	125
The design challenge.....	125
Presentation issues.....	125
Response gathering and marking.....	127
Translating a sample task.....	131
5.5: Conclusions.....	135
New demands on assessment designers.....	135
A critique of GCSE Mathematics.....	136
Chapter 6: Design and evaluation of a prototype eAssessment system	
6.1: Introduction.....	139
6.2: Some general-purpose tools.....	141
The “Printing Calculator”.....	141
The graphing and drawing tool.....	145
6.3: Online marking tools.....	148
6.4: Analysis tools.....	150
6.5: The trial study of GCSE-style tasks.....	152
Aims.....	152
The task set.....	152
Structure of the trials.....	153
Marking.....	154
Data analysis:.....	154
Qualitative analyses:.....	154
6.6: Sample size and composition.....	155
6.7: Results – some case studies.....	158

Table of Contents

6.8: Detailed results for selected tasks	162
Triangle.....	162
Percentages.....	170
Trip.....	174
Taxi Times.....	179
6.9: Effectiveness of the tools.....	182
The printing calculator.....	182
The graph drawing tool.....	183
Marking Tools.....	184
Auto Marking.....	184
6.10: Feedback from pupils.....	187
The value of feedback.....	193
6.11: Practical and technical issues for schools.....	193
Online Delivery.....	193
Other practical issues.....	196
6.12: Conclusions.....	197

Chapter 7: Conclusions

7.1: Introduction.....	199
7.2: The research questions revisited	200
Research question A.....	200
Research question B.....	201
Research question C	202
Research question D	205
7.3: Questions for future research.....	207
Further data collection.....	207
Pupils' mathematical IT skills.....	207
Assessment reform – in any medium.....	208
Formative vs. summative assessment	208
Use of alternative platforms.....	208
7.4: The computer as a medium for “doing mathematics”	210
A new medium.....	210
Computers and writing.....	210
Computers and mathematics.....	211
References.....	213
Appendix A: A prototype online testing system.....	219
Appendix B: Further materials in electronic form	233

Index of Figures

Figure 2.1: Balanced Assessment in Mathematics: a Framework for Balance	15
Figure 2.2: Part of a balancing sheet.....	16
Figure 3.1: A "Balancing sheet" used during the development of World Class Tests	33
Figure 3.2: Floaters - a simulated science experiment	36
Figure 3.3: Sunflower – systematic search for an optimum	37
Figure 3.4: Game of 20.....	38
Figure 3.5: Factor Game – human vs. computer	39

Table of Contents

Figure 3.6:	Queasy - exploring a database.....	40
Figure 3.7:	Water fleas – scientific argument.....	40
Figure 3.8:	Oxygen - exploring multivariate data.....	41
Figure 3.9:	Bean Lab – scientific argument.....	42
Figure 3.10:	Bean Lab - written answer.....	43
Figure 3.11:	Bean Lab – tabulated answer.....	44
Figure 3.12:	Bean Lab - diagrammatic answer.....	45
Figure 4.1:	PIM 6 Equating study- Paper score vs. Digital score.....	63
Figure 4.2:	PIM 7 Equating study - Paper score vs. Digital score.....	63
Figure 4.3:	PIM 6 Equating study - median and quartiles by school.....	65
Figure 4.4:	PIM 7 Equating study - median and quartiles by school.....	65
Figure 4.5:	PIM 6 Equating study: Effect of order of tests on score difference.....	66
Figure 4.6:	PIM 7 Equating study: Effect of order of tests on score difference.....	67
Figure 4.7:	PIM Equating study - digital vs. paper grouped by order of testing.....	68
Figure 4.8:	PIM 6 Equating study - digital vs. paper question facility levels - all schools.....	70
Figure 4.9:	PIM 6 Equating study - digital vs. paper question facility levels - paper test first.....	70
Figure 4.10:	PIM 6 Equating study - digital vs. paper question facility levels - digital test first.....	70
Figure 4.11:	PIM 6 Equating study - paper question facility levels vs. order of testing.....	71
Figure 4.12:	PIM 7 Equating study - digital vs. paper question facility levels - all schools.....	71
Figure 4.13:	PIM 7 Equating study - digital vs. paper question facility levels - paper test first.....	71
Figure 4.14:	PIM 7 Equating study - digital vs. paper question facility levels - digital test first.....	72
Figure 4.15:	PIM 7 Equating study - paper question facility levels vs. order of testing.....	72
Figure 4.16:	PIM 11 Equating study - digital vs. paper question facility levels - all schools.....	72
Figure 4.17:	PIM 6 digital test: Rasch item measures and infit statistic.....	81
Figure 4.18:	PIM 7 digital test: Rasch item measures and infit statistic.....	81
Figure 4.19:	Background screen with irrelevant “mathematical” content.....	87
Figure 4.20:	Colour and text visibility issues.....	88
Figure 4.21:	Digital version of the “number pyramids” task (PIM 8).....	89
Figure 4.22:	The second part of the paper version of the “number pyramid” task.....	89
Figure 4.23:	Constrained responses.....	90
Figure 4.24:	This can be solved by clicking “+” until the two piles look similar (PIM 8).....	90
Figure 4.25:	From "Developing Problem Solving" 8-11.....	92
Figure 4.26:	“Sticks” (PIM 7).....	93
Figure 4.27:	“Nine” (PIM 7).....	94
Figure 4.28:	“Clock (paper)” (PIM 7).....	95
Figure 4.29:	“Clock (digital)” (PIM 7).....	96
Figure 4.30:	Paper and digital versions of “square shapes” (PIM 7).....	99
Figure 4.31:	“Rules” (PIM 7) – digital version.....	101
Figure 5.1:	Relative frequency of answer formats.....	110
Figure 5.2:	Typical style of GCSE question.....	112
Figure 5.3:	Available marks per subtask.....	112
Figure 5.4:	Distribution of marks by topic/activity in the GCSE sample.....	113
Figure 5.5:	Extract from AQA 2003 specimen GCSE papers.....	117
Figure 5.6:	Question from AQA Mathematics Specification A Paper 2, November 2003.....	119
Figure 5.7:	Chi-squared test on data from Figure 5.6 using a free web-based tool.....	119
Figure 5.8:	Probability question from an AQA Specification A GCSE Paper.....	121
Figure 6.1:	The "printing calculator" concept.....	142
Figure 6.2:	Student response using the printing calculator, with working, presented to marker.....	143
Figure 6.3:	Advanced uses of the calculator.....	143
Figure 6.4:	The line-drawing tool.....	146
Figure 6.5:	Drawing a line of best fit.....	146
Figure 6.6:	The marking system.....	149
Figure 6.7.:	Marking a graph question, showing automatic marking.....	149
Figure 6.8:	Example of the data analysis system in action.....	151
Figure 6.9:	Triangle task (paper version).....	163

Table of Contents

Figure 6.10:	Triangle task (C2 variant - computer with rich responses)	164
Figure 6.11:	Triangles mark scheme and sample response (C2 variant)	165
Figure 6.12:	Fully justified answer to part (a) of triangle - from trials of paper test	165
Figure 6.13:	Percentages - paper version	170
Figure 6.14:	Percentages C2 (left) vs. C1 (right) - note different structure	170
Figure 6.15:	Original GCSE markscheme for part (a) of percentages	172
Figure 6.16:	Mark scheme and two examples of marking discrepancies	173
Figure 6.17:	Trip question - paper version	175
Figure 6.18:	Trip task (C2 variant)	176
Figure 6.19:	Trip task (C2 variant) mark scheme and sample answer	178
Figure 6.20:	Taxi times task (rich computer version)	179
Figure 6.21:	Taxi times marking scheme	180
Figure 6.22:	Trial results for taxi times task	181
Figure 6.23:	Pupils successfully engaged with the drawing tool - if not the task itself.	183
Figure 6.24:	Feedback screen (including comment)	187
Figure 6.25:	General attitude vs. Key Stage 3 level (by percentage and number)	188
Figure A.1:	Student registration system	221
Figure A.2:	Logging in	223
Figure A.3:	Taking a test	223
Figure A.4:	The marking system	224
Figure A.5:	Testing automatic marking	225
Figure A.6:	Task statistics display	226
Figure A.7:	Creating a task in Adobe Flash	227
Figure A.8:	Creating a mark scheme in XML	229
Figure A.9:	Schema for the database	230

Index of Tables

Table 3.1:	Scoring Sunflower by inference	37
Table 4.1:	Correlation Coefficients Paper/Digital PIM 6-11	58
Table 4.2:	Key to statistics tables	74
Table 4.3:	PIM6 - Whole equating sample	75
Table 4.4:	PIM6 - Students taking digital test after paper test	76
Table 4.5:	PIM7 - Whole equating sample	77
Table 4.6:	PIM7 - Students taking digital after paper	78
Table 4.7:	PIM 11 - Whole equating sample	79
Table 6.1:	The six test variants used in the trials	153
Table 6.2:	Total numbers of successfully completed & marked school trials	155
Table 6.3:	Students who took both a paper and computer test - composition of sample	157
Table 6.4:	Summary results - whole sample	159
Table 6.5:	Summary results – pupils at KS3 maths levels 4-6	160
Table 6.6:	Computer vs. human marking discrepancies - summary	160
Table 6.7:	Facility levels (average % score) on individual mark scheme points	161
Table 6.8:	Facilities for the individual parts of the "Triangle" task (original marking)	167
Table 6.9:	Percentages of candidates providing working - Triangle Paper vs. C2	167
Table 6.10:	Relative frequencies of responses to the Triangle task	169
Table 6.11:	Percentages part (a) – facilities for comparable parts (whole sample)	171
Table 6.12:	Facilities for the individual parts of the "Trip" task	178
Table 6.13:	Summary of pupil feedback	188
Table 6.14:	Frequent comments	189

1: Outline of the study

...also whats wrong with paper: and manual labour. i dont know what is trying to be proved by using computers, but mine started flashing purple and things and went fuzzy and put me off from answering questions. this WAS NOT HELPFULL you made me very stressed, although it did make me chuckle.

One GCSE student's reaction to a prototype online test (see Chapter 6)

1.1: Rationale

The use of computers to deliver and mark school assessment is likely to grow in the future: in 2004, a speech by the head of the Qualifications and Curriculum Authority for England (QCA) proposed that, by 2009, “all new qualifications” and “most GCSEs, AS and A2 examinations should be available optionally on-screen” (Boston, 2004).

Even though that particular ambition was not realised, continuing interest in computer-based testing, or “eAssessment”, seems likely, because:

- To the government it promises a fast, cheap and objective way of sampling student performance without overburdening (or even involving) teachers
- Examination boards – even as not-for-profit companies – seek to minimise costs, and eAssessment has the potential to eliminate the cost of printing papers and securely transporting them between schools and markers, and to possibly remove the need for manual marking
- To educationalists, eAssessment offers the hope of improving the nature and quality of assessment by increasing the range of task types that can be set in high stakes

1 - Outline of the study

assessment (see e.g. Burkhardt & Pead, 2003) which in turn could influence the taught curriculum

- IT professionals welcome the possibility of large scale funding from government and examination boards

This thesis investigates some of the issues that a large scale move towards computer based tests for both high- and low- stakes assessment might raise in the particular context of mathematics in primary and secondary schools. The work is based on the author's research in connection with a series of computer-based assessment projects conducted between 1999 and 2007.

The term “eAssessment” is used here as a convenient shorthand – it has no formal definition of which the author is aware, but could be used to refer to any assessment which uses information technology for some or all of the following:

- a) Electronically distributing assessment materials to schools
- b) Presenting a task to the candidate on a computer screen
- c) Capturing the candidate's responses to the problem:
This could simply be “the answer”, but for more sophisticated assessment it may be necessary to somehow record what steps they have taken to arrive at that answer
- d) Providing ICT-based resources (such as a data set or simulated experiment) or tools (calculators, spreadsheets) which allow the candidate to progress with an assessment task
- e) Marking the candidates' responses automatically, possibly giving them instant results
- f) Generating unique tests “on demand” by assembling calibrated questions from a bank, using mathematical modelling to ensure consistent difficulty levels. Not having a single annual test paper, which must be kept secure, enables individual candidates to sit tests when their teacher feels that they are ready and, more practically, removes the need for schools to provide enough computers to allow an entire year group to take a test simultaneously
- g) Supporting the more traditional manual marking process (by, for example, allowing markers to work on-screen, possibly remotely)

The maximum logistical and economical advantages are to be expected when most of these aspects are realised, with tests automatically generated on-demand; transmitted over the internet; taken by candidates working on a computer and instantly marked. However, study of

existing mathematics assessments – particularly the GCSE examination in England (see Chapter 5)– suggests that this would require significant changes in the nature of the assessment.

A parallel development in the specific field of mathematics assessment is the observation that current paper-based tests and curricula fail to assess the candidate's ability to combine, adapt and apply their technical skills to solve varied problems – as opposed to simply demonstrating their proficiency at well-practised standard techniques. When this work began, the wording of the National Curriculum in England encouraged such “problem solving” skills to be seen as an optional sub-genre of mathematics (as evidenced by the division of the *World Class Tests* project, discussed in Chapter 3, into separate “problem solving” and “mathematics” strands). More recently, changes to the National Curriculum and initiatives to introduce “functional mathematics” recognise these skills as central to mathematical competence. Experience shows that, unless the high-stakes tests reward such skills they will not receive adequate attention in typical classrooms. Consequently, any near future development in mathematics “eAssessment” will need to address the assessment of “problem solving” skills – yet the type of open questions with free-form answers favoured by, for example, the PISA tasks (*PISA, 2003*) are difficult to reconcile with the requirements for “efficient” computer-based testing discussed above.

1.2: Research Questions

This study will focus on the following research questions:

A. How can eAssessment contribute to the assessment of *problem solving skills* in mathematics?

Computers have the potential to present rich, interactive activities which would not be feasible in traditional tests, thus they have potential for assessment of process skills.

However, designing interfaces and systems to capture, store and mark pupils' responses to such tasks is a challenge, and can be difficult to reconcile with the tightly structured and automatically calibrated nature of typical eAssessment systems. For example, a key “process skill” in a task might be to recognise that arranging the data as a table would help to understand the problem: if the computer presents the student with a fill-in-the-blanks template for such a table, that skill will not be assessed.

Chapter 3 discusses the author's work on the design of the *World Class Tests* as a case study which sought to design, evaluate and deliver (in quantity) computer-based tests of such process skills.

B. What are the effects of transforming an existing paper-based test to computer?

It seems inevitable that, without radical curriculum reform, near-future computer

1 - Outline of the study

based tests will not consist entirely of novel tasks but will rely on many task genres copied from existing assessments. There is some independent evidence that students under-perform on computer-based mathematics assessments, even ones with fairly conservative question types. Is this a “total cognitive load” effect, in which the additional mental effort of operating the computer systematically depresses other aspects of performance, or are the translated tasks now assessing different skills? If so, are these useful, transferrable skills relevant to mathematics or just a matter of knowing how to operate the specific testing software in use? What are the students' and teachers' attitude to computer-based testing? This is of particular concern to a publisher or examination board which offers parallel computer- and paper-based versions of substantially the same test. To this end, the study includes a quantitative and qualitative analysis of a study of a new computer version of an established commercial product – Nelson's *Progress in Maths* tests for younger pupils (Chapter 4). Moving to higher stakes assessment for older pupils we investigate the “state of the art” of GCSE paper tests and examine the issues these raise for the design of comparable computer tests (Chapter 5) and then proceed to design and evaluate parallel computer and paper versions of a prototype test based on traditional GCSE question genres (Chapter 6).

C. How might eAssessment be used to improve the range and balance of the assessed curriculum (and hence, indirectly, the taught curriculum)?

Computers have the potential to enable novel forms of assessment, but, in practice, the economic and logistical incentives noted above are normally achieved through the use of highly structured, multiple choice and short-answer questions. Such “items” are easy to implement and straightforward to mark automatically and can be tightly focussed on a single aspect of performance. They can then be produced in large numbers, individually calibrated, banked and used to automatically construct unique tests with a predictable difficulty profile. How can this be reconciled with current initiatives to better assess process skills and functional mathematics, a field which is reviewed in Chapter 2? What are the implications for GCSE (which, as discussed in Chapter 5 has long eschewed multiple choice in favour of constructed response and, for example, awards substantial credit for showing method)?

D. What do the above issues imply for the technical and pedagogical processes of computer-based assessment design?

What skills and processes are needed to design effective computer based tests, and how do these requirements differ from traditional test design? Are there additional technical requirements for mathematics testing, as distinct from other subjects? What

1 - Outline of the study

practical tools and techniques might help realise the aspirations embodied in the research questions above? These issues pervade this work.

While much use will be made of statistical and psychometric techniques to compare the difficulty and validity of tasks, this study balances this with *design research* (Akker, 2006; Schunn, 2008; Schoenfeld, 2009) and looks in detail at the process of developing and refining new educational products.

Any move to replace an existing test with a computerised version presents an opportunity for change. The analytic design research involved in examining an existing test with a view to computerisation, or an independent attempt at computerising a test, can produce some of the broader issues about the assessed curriculum. Hence, in Chapters 4 and 5, our discussion often extends beyond the practical issues of computerisation to a broader critique of particular tasks. In Chapters 3 and 6 the aim of the design research is to produce and evaluate an improved product.

1.3: Outline of the work

Testing “higher order” or “problem solving” skills

Between 1999 and 2004, the author led the design and evaluation of computer-based items for the problem solving strand of the *World Class Tests* initiative.

The *World Class Tests (WCT)* project, funded by the QCA, ran from 1999 to 2004 and sought to develop two series of tests – one in Mathematics and another in “Problem solving in Mathematics, Science and Technology” – which would identify and challenge gifted and talented students at ages 9 and 13, particularly those whose talents were not being exposed by normal school Mathematics. These tests were – initially – planned to be entirely computer delivered and marked. Critically, although these aims were novel and innovative, this was not a “proof of concept” project: after two years of development, these tests were to be rolled out nationally with a new suite of tests offered four times a year.

A specific requirement of this project was that the items should not simply test students' proficiency in specific mathematical techniques, but should focus on “process skills” - the ability to select, combine and apply these techniques in unfamiliar contexts. This aspect of mathematical performance, often lacking from traditional assessment, has recently been re-emphasised in the National Curriculum (QCA Curriculum Division, 2007).

In the early planning stages of this project, capturing evidence of process skills was recognised as a major challenge, as most conventional approaches to computer-based assessment require the problem to be broken up into discrete steps in order to facilitate

1 - Outline of the study

response capture and automated marking. This often has the side-effect of making the “correct” mathematical technique self-evident to the student (section 5.4 discusses this effect). It also made it difficult to use forms of response such as free-form diagrams and sketches which feature heavily in paper-based tests with similar aims (e.g. Balanced Assessment Project, 1999): although capture (and even automatic marking) of such responses is conceivable, it would require candidates to master a relatively complex user interface within the limited time required by the test, and require the devising of complex marking algorithms for each new question type.

Consequently it was decided that the initial plan for an entirely computer-based test could not adequately assess all dimensions of the domain, and that half of the assessment would be delivered as a conventional paper-based test. In addition, the computer-based test would be augmented by a paper answer booklet. The positive consequence of this was that the software development effort could concentrate on providing rich problem-solving contexts involving games, puzzles, animations and simulated experiments, without being constrained by the problems of capturing pupils' responses. As the designers gained experience, the need for the paper answer book to accompany the computer test diminished, but the separate paper-based test remained essential to ensure a balanced sampling of the domain. The *World Class Tests* project is described in more detail in Chapter 3.

Computerising existing tests

Towards the end of this project, the QCA's ambitions for computerising GCSE and A-Level, mentioned above, were announced (Boston, 2004).

It seemed clear to the author that the ambitious time scale would require the existing repertoire of proven task types to be translated *en masse* to computerised form.

In the light of experience from the *World Class Tests* project, this raised several interesting and important questions, such as:

- Will the change in presentation media alone have an effect on the candidates' performance?
- Current GCSE papers rely on “constructed response” questions which require candidates to show the steps they take to arrive at an answer. How can such questions – particularly those involving mathematical notation – be presented on computer?
- How might such translations affect the assessment objectives and validity of the questions – especially if the translation is performed by computer programmers rather than educational practitioners?

1 - Outline of the study

- Is the current mathematics curriculum still appropriate in a world with pervasive access to information technology, and should curriculum and assessment change to address this? (For instance, GCSE still requires pupils to perform geometric constructions with ruler and compasses).

An opportunity arose to conduct an independent analysis of a computer delivered translation of an established, well calibrated paper-based test. The goal was to determine whether the translation process had affected the test's difficulty, or had changed what was being assessed.

NferNelson's *Progress in Mathematics* is a series of test booklets and scoring guides used by teachers for annual progress monitoring from ages 6-14. It was backed by a large-scale calibration exercise. The publishers had recently produced an online version of the tests consisting predominantly of computer “translations” of the paper questions.

When NferNelson conducted an “equating study” comparing pupils' performance on the paper and electronic versions of the tests at ages 6,7 and 11, they noticed conspicuous variations in the score on specific questions and the suggestion of a general trend towards lower overall scores on the electronic test. The author was commissioned to investigate the issues raised. The work consisted of a re-analysis of the existing equating study data; a critical examination of the design of the tasks and small-scale, detailed observations of the tests in use.

The data from the equating study was examined using multiple techniques – combining visualisation, general statistical tests and Rasch scaling – in an effort to verify the significance of the effects noticed by NferNelson.

The critical analysis of the design of the computer-based tests, drew on the author's experiences on the *World Class Tests* project and earlier experiences in educational software design to identify possible features of the tasks which might have altered their performance compared with the paper originals. For example, simply changing the layout of a multiple-choice question might “draw the eye” to a different answer; or an over-complicated user interface might raise the “total cognitive load” of the question and cause candidates to under-perform on the mathematical aspects of the task.

Additionally, the small-scale trials involved close observation of very small groups of children taking the tests. Since the existing equating study provided substantial statistical data, the objective here was to seek qualitative insights into how children interacted with the individual tasks. Without the need to collect untainted quantitative data, children could work in pairs – to encourage them to vocalise their thinking without prompting by their observers – and observers could intervene when needed. This study is described in detail in Chapter 4.

Analysing GCSE with a view to computer delivery

The nferNelson tests – even in their original paper form – were typified by very short questions with highly structured answers (often either multiple choice, or single numerical answers) which were relatively unchallenging to computerise, yet the study did suggest that pupil performance could easily be affected by design decisions made during the translation process. The *World Class Tests* tasks, in contrast, were free to experiment with the medium without the constraints of comparability with a conventional test, nor did they have to provide coverage across a whole curriculum. Furthermore, each *WCT* task was individually, and thus fairly expensively, custom programmed to the task designer's specifications. If GCSE were to be computerised, it would need an approach midway between the two: the questions would be shorter and more structured than *WCT* but would still need a constructed response element and the ability to capture and, potentially, mark the steps by which candidates arrived at their answers. It should also be possible for an author to rapidly assemble a new task from standard building blocks, with a minimum of task-specific programming.

An analysis of GCSE past papers was conducted to enumerate the types of question which a computer-based test system would need to support and, also, quantify the importance of credit given for working, explanations, partially correct answers or “follow through”¹ which could be challenging to implement in an automatic marking system. Chapter 5 discusses this work, and concludes with a “worked example” comparing alternate approaches to computerising a paper task.

Computerising GCSE mathematics – a design research experiment

It was then decided to construct a prototype system which had the ability to deliver questions adapted from established GCSE genres and allowed the evaluation of possible generic solutions to some of the challenges this faced. In particular, the system would include a “printing calculator” tool (originally devised for, but not used, in *WCT*) designed to capture the steps in calculation-based tasks and a minimal drawing tool for graphs and simple diagrams. Between them, these could enable several genres of traditional questions to be presented. (Other key tools that would be required for a full GCSE implementation would be a way of entering algebraic expressions and a geometric construction tool – these were beyond the scope of the initial study, although they could be incorporated in the future).

The result was a prototype test delivery and marking system which – while insufficiently robust to deliver a “live” high-stakes examination – allowed simple tasks to be constructed

¹ Markers use the term “follow through” where, to avoid penalising a candidate twice for the same mistake, credit is given for applying a correct method to an incorrect previous result.

1 - Outline of the study

rapidly from standard components while still permitting extensive custom programming of experimental task types. Tests were delivered, and data returned, via the internet and responses could be marked manually using a web interface or automatically, using a system of simple rules which could be expanded as necessary with custom algorithms.

Two half-hour paper tests were produced which imitated the style of GCSE papers featuring minor variants of recurring task genres from intermediate-tier GCSE. Two computer versions of each of these tasks were produced – one reduced to the sort of short multiple-choice responses easily implemented on “off-the-shelf” computer-based testing systems, and another using one or more of the “rich” response capture tools. Groups of pupils from local schools (about 270 in total) were each given one of the tests on paper and a computer version of the other test (comprising a mix of the two styles of computer tasks). Pupils also had the opportunity to feed back on the test experience at the end of the computer test.

As well as providing data on comparative performances on the different task types, the study raised interesting issues about the pedagogical and practical/technical issues involved in computer-based tests of mathematics at this level. Practical issues included the readiness of school infrastructure to cope with internet-delivered assessments. Furthermore, the analysis of GCSE papers raised some important questions as to the value of partial credit and method marking as it is currently used in GCSE. Another, unexpected, issue that arose was a hardening of students' attitudes towards computer use in the run-up to high stakes examinations. This experiment is described in Chapter 6.

The author's contribution to this work

The original research work conducted or led by the author, on which this thesis is founded, consists of:

- The analysis of the equating study data in Chapter 4. The equating study itself was designed and constructed by nferNelson
- The design and execution of the analysis and school observations of *Progress In Maths* described in Chapter 4
- The analysis and critique of GCSE tasks in Chapter 5
- The design, development, trialing and evaluation of the prototype eAssessment system described in Chapter 6

Chapters 2 and 3 include the author's commentary on projects in which he has worked as part of a larger team. In the case of *World Class Tests* the tasks shown were conceived by various colleagues at the Universities of Nottingham and Durham, although the author was

1 - Outline of the study

primarily responsible for prototyping, adapting and refining the designs and specifying their implementation to the software developers. The tasks shown in Chapter 4 were designed by nferNelson. The tasks used in Chapter 6 are the author's own paper- and computer-based variants on generic GCSE task types frequently seen on papers by AQA and others, except a few which were the author's adaptations of paper-based tasks from *Balanced Assessment*.

2: Key issues in Mathematics Assessment

A cistern is filled through five canals. Open the first canal and the cistern fills in $\frac{1}{3}$ day; with the second, it fills in 1 day; with the third, in $2\frac{1}{2}$ days; with the fourth, in 3 days, and with the fifth in 5 days. If all the canals are opened, how long will it take to fill the cistern?

*From "Nine Chapters on the Mathematical Art"
- China, approx. 200BC*

2.1: Introduction

This brief review of the field will discuss some issues which are central to the current debate on the future development of mathematics assessment. It will also establish the meaning of some terms and concepts which are used throughout this thesis.

The chapter starts by looking at the "state of the art" of traditional mathematics assessment and the goals of some reform initiatives: a key question for this thesis is how computer-based assessment might help or hinder attainment of these goals.

One overarching issue is the belief that high-stakes assessment tests do not passively measure students' attainment, but also provide the *de facto* specification for much that is taught in the classroom, subverting some of the broader aims of the intended curriculum. One way of addressing this would be to try and reduce the emphasis that society places on assessment and, in particular, discourage the use of aggregated assessment scores as a school accountability measure. Another, possibly more realistic, strategy would be to ensure that the tasks used in high-stakes assessment mirror the activities that we would like to see, every day, in the classroom, while still producing valid accountability data.

2 - Key issues in Mathematics Assessment

To achieve this, it is necessary to ensure that assessments do not consist of a homogenous sequence of short exercises each testing a narrowly designed technical skill, but are “balanced”, comprising a diversity of styles of task which assess the candidate's ability to combine technical proficiency with higher order strategic skills, creativity and insight. One particular area in which current assessments fall short of this is a lack of tasks requiring “problem solving” or “process skills”. Such tasks present particular challenges in terms of design, test calibration and reliable marking.

After setting out the arguments for the above assertions in the context of traditional assessments, this chapter will look at issues specific to eAssessment and consider their relevance to these aspirations.

2.2: The influence of tests on the taught curriculum

High-stakes assessment has a major influence on what is taught in schools. Examination results and summary statistics on test performance play a major role in school accountability, so it is understandable that teachers concentrate on those aspects of mathematics which appear in the tests:

...although most secondary teachers recognised the importance of pedagogic skills in mathematics, they often commented on the pressures of external assessments on them and their pupils. Feeling constrained by these pressures and by time, many concentrated on approaches they believed prepared pupils for tests and examinations, in effect, ‘teaching to the test’. This practice is widespread and is a significant barrier to improvement.

Mathematics – Understanding the Score (Ofsted, 2008).

Not only do teachers concentrate on the **topics** covered by the test, but they model their classroom activities on the **format** of the test (Shepard, 1989). So, if a test consists of short questions each focussed on a specific mathematical technique (such as factorising a quadratic) the daily work in most classrooms will be dominated by similar, short exercises.

In Australia, a study of the classroom effects of the introduction of a novel final examination (Barnes, Clarke, & Stephens, 2000) found evidence of mandatory assessment driving teaching practices throughout the secondary school, supporting their observations that:

- “attempts at curriculum reform are likely to be futile unless accompanied by matching assessment reform” and
- “assessment can be the engine of curriculum reform, or the principal impediment to its implementation”.

2 - Key issues in Mathematics Assessment

The influence of assessment on the taught curriculum was also recognised in the report *Mathematics Counts* (Cockroft, 1982) - a key influence on the design of the UK national curriculum. Here, the particular concern was the indirect impact of assessments, which had previously served as entrance requirements for mathematical subjects at university, on courses of study for lower attaining students. Despite this recognition, the implementation of the National Curriculum has still been strongly influenced by assessment:

“...most of the available resources, and public and political attention, have been concentrated on the tests which are given at the end of the Key Stages to yield overall levels or grades...”

Inside the Black Box (Black & Wiliam, 1998)

The studies mentioned above strongly suggest that, whatever the *intended curriculum* and regardless of the stated standards for “best practice” in the classroom, it is the contents of high-stakes tests which will have the dominating influence on the *taught curriculum* in typical schools. The implication is that designers of assessment should pay as much attention to the formative properties and pedagogical validity of their tasks as to their psychometric properties, even for a summative, end-of-course examination.

2.3: Achieving “balance” in assessment – a case study

“An assessment which focusses on computation only is out of balance. So is one that focusses on patterns, functions and algebra to the exclusion of geometry, shape and space, or that ignores or gives a cursory nod towards statistics and probability. Likewise, assessments that do not provide students with an opportunity to show how they can reason or communicate mathematically are unbalanced. These are content and process dimensions of balance, but there are many others – length of task, whether tasks are pure or applied and so on.”

From: Balanced Assessment for the Mathematics Curriculum
(Balanced Assessment, 1999)

Given the potential of assessment to distort the taught curriculum, it would seem desirable to ensure that assessments comprise a “balanced diet” of diverse mathematical tasks which combine curriculum knowledge and technical skills with higher-order “strategic”, “process” or “problem solving” skills. Can such a test be delivered and reliably marked, and what techniques could be used to ensure that the test is, indeed, “balanced”?

Balanced Assessment in Mathematics is an ongoing series of related projects² aimed at producing better-balanced alternatives to the conservative assessments commonly used in the

² The Balanced Assessment projects represent a collaboration between groups at the University of Nottingham, University of California at Berkeley, Michigan State University, Harvard University with funders including National Science Foundation, McGraw Hill and The Noyce Foundation.

2 - Key issues in Mathematics Assessment

USA. While not part of the research presented in this thesis, it represents an important background source for this work.

All US states require that their mathematics programmes are “aligned” with state standards. Many of these are influenced by the standards developed by the USA National Council of Teachers of Mathematics, published as *Principles and Standards for School Mathematics* (NCTM, 2000). These include principles for assessment which recognise the importance of a wide range of assessment techniques “including open-ended questions, constructed response tasks, selected response items, observations, conversations, journals and portfolios” and observe that “Constructed-response or performance tasks may better illuminate students' capacity to apply mathematics in complex or new situations”

Despite this, partly because of cost pressures, the high-stakes tests in these states often rely on batteries of short, multiple choice, questions. The state-mandated mathematics tests from the California Standardised Testing and Reporting (STAR) programme (California, 2008) are one example of this.

In contrast, the aim of the Balanced Assessment project is to produce assessments aligned with the NCTM *Standards* (NCTM, 1989) which included test items and packages covering a multi-dimensional domain of task types, content knowledge, and process skills.

The project devised a “framework for balance” (Figure 2.1) which allows tasks to be classified against multiple dimensions of content, process, contexts and task types.

Dimensions of Balance

Mathematical Content Dimension

- **Mathematical content** in each task will include some of:

Number and Operations including: number concepts, representations relationships and number systems; operations; computation and estimation.

Algebra including: patterns and generalization, relations and functions; functional relationships (including ratio and proportion); verbal, graphical tabular representation; symbolic representation; modeling and change.

Measurement including: measurable attributes and units; techniques and formulas.

Data Analysis and Probability including: formulating questions, collecting, organizing, representing and displaying relevant data; statistical methods; inference and prediction; probability concepts and models.

Geometry including: shape, properties of shapes, relationships; spatial representation, location and movement; transformation and symmetry; visualization, spatial reasoning and modeling to solve problems.

Mathematical Process Dimension

- **Phases** of problem solving include some or all of:

Modeling and Formulating;
Transforming and Manipulating;
Inferring and Drawing Conclusions;
Checking and Evaluating;
Reporting.

- **Processes** of problem solving, reasoning and proof, representation, connections and communication, together with the above phases will all be sampled.

Task Type Dimensions

- **Task Type** will be one of: design; plan; evaluation and recommendation; review and critique; non-routine problem; open investigation; re-presentation of information; practical estimation; definition of concept; technical exercise.
- **Non-routineness** in: context; mathematical aspects or results; mathematical connections.
- **Openness** –tasks may be: closed; open middle; open end with open questions.
- **Type of Goal** is one of: pure mathematics; illustrative application of the mathematics; applied power over a practical situation.
- **Reasoning Length** is the expected time for the longest section of the task.

Circumstances of Performance Dimensions

- **Task Length**: in these tests most tasks are in the range 5 to 15 minutes, supplemented with some short routine exercise items.
- **Modes of Presentation, Working and Response**: these tests will be written.

Figure 2.1: *Balanced Assessment in Mathematics: a Framework for Balance*

2 - Key issues in Mathematics Assessment

Task Name	Task total	weighted mins'	Avgg min/ast	Tasks >>>>	Choc'it	Drkr	Sort	House	Check an	Design	25	Kilney	Falagng	Cross the	hecklr	Task #
Task #					polyhdr	a cab	them	in.hurry	blometel	a text	blowup	stones	poda	ltd-box	game	access
Task length	525	525	48	45	45	45	45	45	45	45	45	45	60	60	45	Task length
Weight factor	11	11	1.00	1	1	1	1	1	1	1	1	1	1	1	1	Weight factor
Strategic aspects																
Task type	Open investigation	2	105	10	1										1	Open investigation
	Non-routine problem	3	135	12			1		1	1						Non-routine problem
	Design	3	150	14						1			1		1	Design
	Plan	1	45	4					1							Plan
	Recommendation	1	45	4		1										Recommendation
	Review/critique	0	0	0												Review/critique
	Re-presentation	0	0	0												Re-presentation
	Definition	0	0	0												Definition
	Exercise	1	45	4								1				Exercise
	Other	0	0	0												Other
Non-routine:	Context	9	435	40	1	1		1	1	1	1		1	1	1	Context
	Math results/aspects	2	90	8	1				1							Math results/aspects
	Math connections	2	105	10			1						1			Math connections
Openness	Open-end	4	195	18	1			1							1	Open-end
	Open-middle	5	240	22		1	1			1	1			1	1	Open-middle
	Reasoning length	265	12600	24	15	15	10	45	10	25	45	10	30	15	45	Reasoning length
Goal type	Pure mathematics	2	90	8	1		1									Pure mathematics
	Illustrative application	2	90	8					1		1					Illustrative appln
	Applied power	7	345	37		1		1	1	1		1	1	1	1	Applied power
Context type	Student life	6	285	26		1			1	1	1			1	1	Student life
	Adult life	3	150	14					1				1	1		Adult life
	Curriculum	0	0	0												Curriculum
	Mathematics	2	90	8	1		1									Mathematics
Phases	Formulation	35	168	15	7	3	2	3	2	3	2	2	4	3	4	Formulation
	Transformation	42	204	19	1		4	1	8	4	6	6	6	4	2	Transformation
	Interpretation	16	77	7	1	2	4	2			2	2		3		Interpretation
	Evaluation	10	45	4	1	3	2		2						2	Evaluation
	Communication-rprt	6	27	2		2			2						2	Communication-rprt
Content aspects																
Number and Quantity		13	62	6												Number and Quantity
Algebra and Function		20	90	8												Algebra and Function
Geometry, space and shape		37	179	16	10		10		10	8	4		8		7	Geom. space and shape
Data, statistics and probability		30	150	14		10						10		10		Data, stats, prob
Other mathematics		10	45	4				10								Other mathematics
Content - Level 2																
Number and Quantity	Concepts and repts	0	0	0												Concepts and repts
	Computation	4	195	18					1	1		1		1		Computation
	Estimation/measurmnt	4	195	18					1	1		1		1		Estimn/measrmt
	Number theory	0	0	0												Number theory
	&props of number															&props of number
Algebra and Function	Patterns and generalizt	0	0	0												Patterns and generalizt
	Functional relationshi	2	90	8			1		1							Functional relationshi
	Graphical/tablr reprn	2	90	8			1		1							Graphical/tablr reprn
	Symbolic reprn	1	45	4			1									Symbolic reprn
	Forming/solving relns	1	45	4					1							Forming/solving relns
Geometry, space and shape	Properties of shapes	5	240	22	1				1	1		1		1		Properties of shapes
	Visualization and repr	4	195	18	1				1			1		1		Visualization and repr
	Location and movemnt	1	45	4											1	Location/movement
	Transfmn/symmetry	2	105	10						1		1				Transfmn/symmetry
	Trigonometry	2	105	10					1			1				Trigonometry
Data, statistics and probability	Concepts	1	45	4		1										Concepts
	Data collectn/analysis	3	150	14		1						1		1		Data collectn/analysis
	Probability models	2	105	10								1		1		Probability models
	Simulation	0	0	0												Simulation
Other mathematics	Discrete mathematics	1	45	4				1								Discrete mathematics
	Pre-calculus	0	0	0												Pre-calculus
	Math structures	0	0	0												Math structures
Scoring aspects																
Scoring categories	Probsolv/reasoning	0	0	0												Probsolv/reasoning
	Communication	0	0	0												Communication
	Number	0	0	0												Number
	Algebra	0	0	0												Algebra
	Geometry	0	0	0												Geometry
	Data	0	0	0												Data
	Other	0	0	0												Other
Calculation zone																
11																
Total # ftl weight/average																
					1	1	1	1	1	1	1	1	1	1	1	

Figure 2.2: Part of a balancing sheet

2 - Key issues in Mathematics Assessment

As well as guiding the work of the task designers, the framework provided a systematic way of ensuring the validity of tests. Each task was evaluated against the framework by assigning a weight to its contribution to each dimension. These weights could be normalised and summed across a test, to ensure that the test as a whole presented a balanced range of content, process types and task styles. Figure 2.2 Shows an excerpt from such a balancing sheet.

This approach allows for the construction of tests from substantial questions, each of which contributes to several dimensions of the domain. It also makes it easier for a proportion of the tasks to be developed using a “context-led” approach in which, rather than constructing each task around a particular statement in the curriculum specification, the designer explores an interesting context, allowing the content and processes to be assessed to emerge naturally.

Marking Balanced Assessment

The original *Balanced Assessment* packages, intended for classroom use, avoided the use of standard point-by-point mark schemes, which identify the anticipated steps in the solution and assign marks for the presence of correct answers and/or evidence of correct working for each step. Instead, they adopted a “holistic” system in which the pupil's response to each task was graded on a generic 4-point level scheme (“The student needs significant instruction”, “the student needs some instruction”, “the student's work needs to be revised”, “the student's work meets the essential demands of the task”). Markers were given short statements characterising typical performances (on the task as a whole) at each level, each illustrated by specimen student work representative of that level.

This system was originally chosen because it was well suited to more open-ended, unstructured problem types which could be solved in multiple ways or had multiple “correct” answers. However, when the tasks were used for more formal assessments requiring defensible, standardised scoring, it was found that the system placed unsustainable demands on markers, who had to familiarise themselves with large volumes of sample work through lengthy training procedures, conferring on and resolving “borderline” cases.

The Balanced Assessment tests (MARS, 2000) used more conventional mark schemes which allocated marks to specific responses or techniques, although these were somewhat more detailed than those used in (for example) GCSE. Some of the techniques and terminology of holistic scoring continued to play an important role in the development of mark schemes and when determining “grade boundaries” for tests. During the trials of these tests, the markers worked closely with the task designers on revising the mark schemes to ensure that they properly reflected students' performance and could be reliably used by other markers. Specimen student work, selected from the trials, continued to play an important role in scorer training and standard setting for subsequent Balanced Assessment tests.

2.4: Assessing Problem Solving, Functional Mathematics and Process Skills

There is a general recognition in the mathematics education community that “problem solving” is an important mathematical activity, and that many students gaining good mathematics qualifications still lack “functional mathematics ³” ability. This may be partly due to a deficiency in basic skills but, more importantly, may indicate lack of ability to apply skills successfully learnt in the mathematics classroom to real-world situations. For example, one study showed that a group of extremely able “A Level” mathematics students, although expecting top scores on the algebra-heavy exam, consistently failed to use algebra to solve a series of planning and optimisation tasks for which it would have been ideal (Treilibs, Lowe, & Burkhardt, 1980).

Although the ability to use and apply mathematics, including problem solving, has always been part of the envisaged National Curriculum, the implementation of the curriculum has, in practice, focussed on “Attainment Targets” - descriptions of specific aspects of mathematical performance under broad headings such as “Number and algebra”, each composed of statements such as:

“(Pupils) recognise approximate proportions of a whole and use simple fractions and percentages to describe these”.

*National Curriculum for Mathematics, Key Stage 3
(QCA Curriculum Division, 2007).*

These targets were used as the framework for all teaching and assessment, encouraging each aspect to be considered in isolation. “Using and Applying Mathematics” comprised a separate Attainment Target, encouraging it to be taught and tested separately – if at all.

“That ‘using and applying’ should have been an aspect of study in each content area, and that tasks including content from more than one area should have been included was recognised: it was not taken seriously because it did not fit the model”

Problem Solving in the United Kingdom (Burkhardt & Bell, 2007).

In response to concerns over this bias, the 2007 revision of the National Curriculum for Mathematics in England places a new emphasis on the “key processes” of:

- Representing
- Analysing
- Interpreting and Evaluating
- Communicating and Reflecting

3 Functional Mathematics is the term currently in vogue in the UK, as a complement to “Functional (il)Literacy” - other terms in use worldwide include “mathematical literacy” or “quantitative literacy”.

2 - Key issues in Mathematics Assessment

According to the National Curriculum these are “*clearly related to the different stages of problem-solving and the handling data cycle*”. However, despite these changes at the top, descriptive level of the curriculum document, in the detailed specification “using and applying” is still listed as a separate “attainment target”.

As an example, part of the curriculum's definition of “Analysing” is that a pupil should:

“appreciate that there are a number of different techniques that can be used to analyse a situation”

It seems self-evident that a child's proficiency in the well defined techniques described by the earlier “statement of attainment” on fractions is more amenable to reliable, quantitative assessment than making inferences of their “appreciation” of a holistic characteristic of mathematical problems. So what type of task can assess such problem solving skills?

According to some widely accepted interpretations (e.g. Bell, 2003; PISA, 2003; Steen, 2000) common features of problem solving and functional mathematics tasks include:

- *Non-routine tasks* – the task is not an obvious variant on a generic template that a student might have been drilled on; the mathematical techniques required are not immediately suggested by the form of the question
- *Extended chains of reasoning* – the student must autonomously perform several steps, combining mathematical techniques, to arrive at the answer, without being led through the process by step-by-step instructions, sub-questions or *pro-forma* response templates
- *Focus on analytic reasoning* – rather than recall of imitative techniques
- *A balance of task-types* – such as *design/plan, optimise/select, review/critique, model...* not simply “solve” or “compute” as in many traditional exercises.
- *Realistic contexts* – which represent plausible applications of the mathematics in a range of settings.

These requirements conflict with some of the practical pressures on those designing and delivering high-stakes, large scale assessment, such as:

- *Defensibility and consistency* – a task designed to be “non-routine” could be accused of being outside the syllabus or too difficult in a system where schools and candidates have come to expect minor variations on well-established task types. Methods need to be established for regulating the difficulty of tests – one solution being trialling of all new tasks and tests to collect calibration data and identify unsuitable tasks. This is

2 - Key issues in Mathematics Assessment

common in assessment design, but is not standard practice in key high-stakes tests such as GCSE and A-Level.

- *Curriculum coverage* – Since attempting to exhaustively test every aspect of the curriculum at each sitting would be impractical, examinations in most subjects only “sample” the curriculum (so, an English Literature examination would not expect the student to answer a question on every aspect of every set book). Mathematics examinations are unusual in the extent to which they aim to “cover” the entire syllabus, only resorting to sampling at the finest level of detail. For example, a typical GCSE examination will always have questions on every key topic, with the only variations being (for example) whether the trigonometry function is Sine, Cosine or Tangent or whether the descriptive statistic to be calculated from data is the mean, median or mode. The need to cover so many topics in a few hours of testing inevitably limits the depth to which understanding can be assessed.
- *Mark allocation* – another expectation of existing assessments is that each mark should be easily attributable to evidence of a particular skill. There is evidence at GCSE of questions being deliberately fragmented to facilitate this by, for example, splitting “solve this quadratic equation” into “(a) factorise this quadratic; (b) now solve the resulting equation”. In England:

“The examination boards were instructed to assess the statements of attainment, with a specified number of marks on a test for each level – each score point must directly relate to one statement of attainment”

(Burkhardt & Bell, 2007)

- *Economy* – it is soon clear from any observation of past papers that similar questions are used year after year with minor variations in the numbers used or the surrounding context. One great saving of this approach is that the difficulty of such tasks can be reasonably assumed to be consistent (especially if candidates have learnt to spot “task types” and have rehearsed the solutions) so it is not necessary to run calibration trials of each new sitting. Another economically prudent technique is to divide the syllabus up between a team of examiners, allowing each one to work independently, whereas writing richer tasks which assess several topics in combination would require more co-ordination of the team.

These constraints are most easily satisfied by short test items which test each curriculum statement in isolation, rather than extended tasks drawing on several concepts. They favour a design process which takes each syllabus statement and constructs a mathematical exercise to test that statement, making it easy to verify coverage of the syllabus. They encourage

questions with simple right-or-wrong answers, rather than tasks which elicit and identify varying levels of mathematical performance. These tendencies are clearly at odds with the characteristics of problem solving tasks proposed earlier.

Problem solving vs. “word problems”

A distinction should be drawn between a valid “problem solving” task and a mathematical exercise presented as a narrative description (often described as a “word problem”).

When contexts are constructed around syllabus statements, the results are often clearly contrived and bear no resemblance to real life (“Maria is thinking of a number”). In other cases, the application is superficially plausible, but makes indefensible simplifying assumptions (for example, the currency exchange example in section 5.3 is mathematically correct but does not represent what really happens at a currency exchange desk). In extreme cases, the choice of context actually undermines the validity of the mathematics (such as the “seeds in a pot” example in section 5.3 which embodies a significant mathematical fallacy).

Presenting a pure mathematical problem as a “word problem” usually adds a comprehension step, potentially increasing the cognitive load, and hence the difficulty of the task. However, this does not automatically add a requirement for strategy or insight if the underlying task is still a routine exercise. There is a tendency, in the pursuit of fairness and consistency in assessment, for the language used in word problems to become codified, following strict rules which students are taught along with the mathematics and thus losing any connection to the real world application of mathematics.

By contrast, a good problem solving task represents an authentic problem, in a carefully chosen context, that someone might actually encounter in the real world. In an easy task, any added difficulty of the “comprehension step” might be justified by an interesting, relevant context that helps the student engage with the mathematics. In a harder task, the difficulty should come from a genuine requirement for insight or strategy, not by decoding the question to find the “right sum to do”.

Assessments with strong problem solving elements have been produced: PISA, *World Class Tests* and *Balanced Assessment in Mathematics* are examples. They are rarely used for “high stakes” assessments, although PISA, the most widely known and used example, is widely used as a benchmark for comparing international standards (PISA, 2003). These assessments demonstrate that, provided the practical pressures are balanced with a firm commitment to meeting the criteria for task quality, viable large-scale assessments of problem solving skills can be produced.

Note: as the terms “problem solving”, “strategic skills”, “functional skills”, “process skills” and similar concepts have overlapping definitions which vary between different projects and contexts, the remainder of this thesis will use “problem solving” as an umbrella term.

2.5: The challenges of computerisation

Is mathematics “harder” on computer?

How sensitive are mathematics test questions to the sorts of design changes required by computer-based assessment, and is it reasonable to expect a computer-based test to be equivalent to the traditional test from which it was adapted?

A student who is slow at typing, or unfamiliar with the mouse will obviously be disadvantaged when taking a timed test. In addition to the direct time penalty, the role of “cognitive load” in the effectiveness of learning materials has been widely researched (see e.g. Sweller, 1994) showing that instructional materials or environments that make heavy demands on “working memory” can result in less effective learning, so it is not surprising that the “extraneous cognitive load” of taking a test on computer might further impact the performance of a student with poor IT skills.

However, another study by Russel (1999) compared performance on paper and computer questions requiring typed, textual answers and found that, although performance in all subjects was correlated with keyboard skills:

“for math tests, performance on computer underestimates students' achievement regardless of their level of keyboarding speed.”

(Russel, 1999)

The latter was in contrast to his findings that, with adequate typing skills, subjects performed better on open-ended language arts subjects.

Russel's computer proficiency tests focussed on keyboarding skills and the maths tests chose questions that could be answered with a short, text answer. Russel notes:

“...despite efforts to include items that did not require students to draw pictures or graphs to receive credit ... about 20% of the students who performed the math test on computer indicated that they had difficulty showing their work and/or needed scrap paper to work out their solutions.”

(Russel, 1999)

Other studies in the US have suggested that, whereas multiple choice questions perform similarly on computer and paper, “constructed response answers” appear harder when presented on computer. The NAEP study (Sandene et al., 2005), which compared paper- and

2 - Key issues in Mathematics Assessment

computer-based tests at US grades 4 and 8 found that lack of computer proficiency “may introduce irrelevant variance into NAEP on-line mathematics test performance”.

At grade 8, NAEP used 16 multiple choice questions, 8 “short constructed response” answers (“which required such actions as entering a number or clicking on a line segment”) and 2 “extended constructed response” questions (“which asked the student to provide an answer and enter an explanation”) and noted that the observed discrepancies between paper and computer were, on average, about twice as large for constructed-response questions as for multiple-choice.

The constructed response issue was correlated with the need to “considerably change” the presentation for computer delivery. However, apart from noting that three of the four items requiring “considerable change” involved the need to enter decimals, fractions and mixed numbers, no details were given of the criteria used to decide whether the changes had been “considerable” or “minimal”.

Capturing working and “method marks”

The US studies discussed above are not a particularly good fit for UK mathematics examinations such as GCSE. Whereas multiple-choice questions are widely used in US mathematics assessment – over half of the NAEP questions were multiple choice – they are rarely used at mathematics GCSE, where the predominant “mode of response” here is a short numerical or algebraic answer.

Furthermore, GCSE mark schemes attach considerable importance to “method marks” - awarded for the sight of correct technique or knowledge in the student's working, even in the absence of a correct final answer (see Section 5.4). This was not discussed by the NAEP or Russel studies. Consequently, a large proportion of GCSE tasks require the student to supply an answer **and** show their method, meaning that they most closely resemble the “Extended constructed response” tasks, only three of which were used in the NAEP study.

If the practice of awarding marks for “showing your work” is to be maintained, will it remain valid or does it introduce a new requirement for the student to “present” their work to the computer adding extraneous cognitive load to the problem? Is the practice of sufficient value to justify the complication of continuing it?

A natural medium for “doing mathematics”?

So, is the computer a “natural medium for doing mathematics” for students in primary and secondary education, or does it add an extra cognitive load to an already difficult subject?

2 - Key issues in Mathematics Assessment

Generic communication and reference tools – such as word processing, email and online or CD information resources – are broadly applicable to most subject areas, but are not directly suited to presenting “the language of mathematics”. Entering an expression such as

$$\sqrt{5^2+12^2} \quad \text{or} \quad \frac{22}{7}$$

in a word processor requires the use of a specialist “equation editor” module, that may not be installed as standard and is primarily designed for the reproduction of a predetermined expression rather than as a tool for working mathematically. The most ubiquitous “generic” mathematical tools are probably the spreadsheet and the on-screen “calculator” simulation. The principle use of spreadsheets in primary and secondary education is data handling – their more sophisticated applications in modelling and discrete mathematics fall outside of most assessed curricula, and so the students' fluency with them should not be overestimated. Anecdotally, the author has found that some mathematics teachers regard the use of spreadsheets as an issue for the separately taught ICT curriculum⁴. Later in this study, the use of an on-screen calculator as a familiar medium for capturing students' working is discussed.

Tools for graphing of algebraic functions (e.g. Butler, & Hatsell, 2007; Pead, 1995) interactive geometry (“Cabri,” 2009; “Geogebra,” 2009; see also Schwartz, Yerushalmy, & Wilson, 1993) and algebraic analysis (“Mathematica,” 2009; “Maxima,” 2008)- all require significant training or experience to use effectively, and while the graphing and geometry tools mentioned are partly designed for secondary education, the computer algebra systems are more suited for university level. Clearly such tools could only be used in assessment if students were required to learn to use them as part of the curriculum.

There may be a qualitative difference between the software tools useful for mathematics and those used in other areas of study. The familiar word processor can aid writing and presentation independently of the subject being examined, it is “content free” – even disabling the spelling and grammar checker, should these subvert the assessment goals, does not fundamentally change the nature or operation of the main tool. In contrast, authentic mathematical applications are likely to embody and automate the very techniques and concepts being tested, providing instant answers to many current question types.

A computer based test might, therefore, require purpose-designed tools, which allowed the entry and manual manipulation of mathematical notation and diagrams without providing any inappropriate help to the candidate. Can these be made sufficiently easy to use so as to not present an extraneous cognitive load, or would they need to be taught along with the subject? Such tools would only make sense in an examination context, so learning their use might not

4 This is discussed in the Professional Development module “ICT: Making effective use of resources” in *Bowland Maths* (Swan, Pead, Crust, & H Burkhardt, 2008)

equip students with any useful, transferrable skills for the future. Is this a fundamental problem, or is it an indication that the current curriculum focusses on mechanical skills that have been devalued by computer technology? Chapter 6 describes a design research study of a few possible approaches to this.

Automatic marking

At a superficial level, it would seem that marking the formal language of mathematics should be less challenging than coping with the complexities of natural language – a problem that commercially-available systems such as *C-rater* (Leacock & Chodorow, 2003) or *QuestionMark* (Intelligent Assessment, 2006) claim to have reliably solved⁵.

For questions that naturally produce a short numerical or multiple-choice answer, marking is easily reduced to simple rules which can easily be automated. More complex responses – such as graphs and charts; problems with multiple correct answers; algebraic expressions or answers which depend on previous results can, at least in theory, be analysed and scored algorithmically, but this makes three major assumptions:

- That the test delivery system used is fully programmable, and doesn't simply offer a range of templates for marking simple numerical, textual or multiple choice responses.
- That the time, skills and resources are available to devise, implement and thoroughly test algorithmic marking methods for each new problem type encountered.
- That the response can be captured by the computer in a consistent, unambiguous form without making unrealistic demands on the IT skills of the candidate during the test, potentially distracting them from the actual task.

While most types of response can, in principle, be captured and marked, will the practical effect be to restrict tests to a small repertoire of task types for which proven capture and marking algorithms are available? This would unbalance the assessment and thus distort the implemented curriculum.

2.6: The implications of eAssessment for assessment reform

It is clear from the design issues raised above that the easiest and cheapest approach to eAssessment is to use questions with multiple choice or simple numeric answers that require minimal typing. It is also possible to use visually richer interactions, asking candidates to “drag and drop” tiles containing text or symbols to their correct positions, but these are still a

⁵ Testing the claimed efficacy of these products in a mathematical context would be interesting, but is beyond the scope of this work, which will assume that marking of short text-only answers is a soluble issue.

2 - Key issues in Mathematics Assessment

highly constrained mode of response – effectively variations, albeit potentially useful – of multiple choice.

Such short-answer questions are particularly convenient in the context of “on demand” testing where individual students can elect to take the test at any time they choose. Since this precludes the traditional practice of keeping the current test paper secret, each student must be given a “unique” test automatically compiled from a large bank of questions.

Statistical techniques such as Item-Response Theory (IRT) can be used to reliably predict the score distribution and ensure that tests are of consistent difficulty. The issue here is that IRT is predicated on each “item” measuring a single, well defined aspect of the subject with a quantifiable difficulty. IRT-related techniques such as Rasch Scaling (Bond & Fox, 2001) have been successfully applied to extended tasks by careful mapping of “items” to individual elements of performance described in the mark scheme, as has been done with selected *World Class Test* tasks (Ridgway, Nicholson, & McCusker, 2006). However, using IRT to programatically build tests, rather than analyse them, is only straightforward when each task represents a single, independent item addressing a particular point on the syllabus, favouring a short answer or multiple-choice test.

Another factor is that each question – and the rules by which it is to be scored – must be implemented in software. From an economic point of view it is highly desirable to reduce this to a routine data-entry task, rather than requiring a skilled programmer to code each question and devise algorithmic rules for identifying the correct response. Again, this favours short questions with a single, clearly identifiable right answer.

These requirements – large banks of short questions, each targeted on a single curriculum statement, with simple right-or-wrong numerical or multiple-choice responses – seem difficult to reconcile with the aspirations for assessment discussed earlier. Even traditional “constructed response” tests such as GCSE would need significant changes to comply with these constraints.

Computers can, in principle, present rich, interactive problems with multimedia, simulated experiments and new mathematical tools – but is it feasible to produce viable replacements for current high-stakes assessments which exploit this potential? That is the focus of this study.

2.7: A note on “formative assessment”

In addition to raising issues about the influence of tests on teaching, the work by Black and Wiliam (1998) produced compelling evidence for the efficacy of *formative assessment* techniques which avoid summative scores in favour of rich, diagnostic feedback to pupils. These techniques rely on a substantial change in the day-to-day teaching style of most teachers, and the gains found by these studies cannot be reproduced by better test questions alone. Black has since noted that the ideas behind formative assessment have been misappropriated to mean “regimes of frequent summative testing” (Black, 2008)

This chapter has presented some arguments for the inclusion of richer, more open tasks in high stakes assessment, but as long as such tasks are used in a summative fashion there is no evidence to suggest that they will lead to the specific gains noted by Black & Wiliam. In fact, the same evidence suggests that introducing any summative element (such as giving pupils scores or grades) distracted pupils from the rich feedback and risked negating any gains.

However, the influence of high-stakes tests on both the content and style of classroom teaching has been noted above. If these tests are composed of rich, open tasks, then it is easier for teachers to justify devoting classroom time to the formative use of similar tasks or past test questions. Hence, there is a strong argument for ensuring that any assessment task has the potential to provide a formative experience, even where it is to be used in a summative way.

In addition – if computers are to be used in formative assessment – the ability to capture pupils' working and richer forms of response could enable teachers to provide formative feedback on pupils' work, whereas systems which simply record the answer offer little choice beyond a right/wrong mark. One issue encountered during the *Progress in Maths* evaluation (Chapter 4) and some stages of the *World Class Tests* development (Chapter 3) was the difficulty in reviewing pupils' actual work after they had completed the test. A key design feature of the prototype system used in Chapter 6 was the ability to visually reproduce pupils' work for the markers.

In contrast, a popular feature of many computer-based testing systems is their ability to produce impressive summary reports and standardised scores, which is certainly not conducive to the type of formative assessment envisaged by Black & Wiliam.

Overall, although some aspects of this work may have direct or indirect implications for formative assessment, the focus here is on summative assessment.

3: Assessing problem solving skills: a case study

I love deadlines. I like the whooshing sound they make as they fly by.

- Douglas Adams

3.1: Introduction

The first research question posed in this thesis was “How can eAssessment contribute to the assessment of problem solving skills” and the importance of such skills in any balanced assessment of mathematics was discussed in Chapter 2. It has already been suggested that some styles of computer-based assessment might actually make it harder to assess problem solving, by favouring short or highly structured questions with simply expressed answers. This chapter considers how the principled design of interactive, computer-delivered tasks can enable the assessment of problem solving and process skills in ways that would not be possible in a conventional test.

The subject of this chapter is a case study of a project which specifically focussed on problem solving skills, without the usual obligation to assess the wider mathematics curriculum. This provides a contrast to later chapters which take a more pragmatic view and consider the issues of replacing more conventional assessments of the established curriculum. It should, however, be noted that this project **was** required, after an initial research phase, to deliver, in quantity, externally marked assessments which were published and administered by an awarding body. This places it in a slightly unusual position between pure “insight” research projects, which might study a few tasks in great detail, and regular assessment production.

The author was the lead designer for the project strand working on computer-based problem solving tasks.

3.2: The *World Class Tests* project

The brief

The *World Class Tests* were the central part of the QCA/DfES funded *World Class Arena* programme, intended to provide support for “gifted and talented students”. A particular focus was to identify, engage and challenge those students whose ability might not be apparent from their performance on day-to-day classroom activities (so-called “submerged talent”).

The product, as originally conceived by the Government in 1999, would consist of computer-delivered assessment tests for students at ages 9 and 13, with four new test sittings available every year. Early in the tendering process, this was altered to include a mix of computer- and paper- based tests, for reasons discussed later.

There would be two separate sets of tests: “Mathematics” and “Problem solving in mathematics, science and technology”. This chapter concentrates on the development of computer-based tasks for the “problem solving” strand and the issues arising from this process.

Educational principles

Although aimed at more able students, a key constraint of the design, which has resonance with some models of functional mathematics/mathematical literacy, was that the tasks should **not** require above-average curriculum knowledge, but should focus on more sophisticated reasoning and insight (see e.g. Steen, 2000). It was therefore necessary to agree on a clear description of these “process skills” and methods for ensuring that each test covered this domain adequately. Although there was no strictly defined body of content knowledge which had to be assessed, each test sitting was expected to include a range of topics covering mathematics, science and technology. The chosen solution was a development of the “framework for balance” model devised by the *Balanced Assessment project*.

For the *World Class Tests* this was adapted to produce a “Domain framework in mathematics and problem solving” (Bell, 2003). The definitions of problem solving adopted by the OECD PISA assessments (PISA, 2003) were also referenced for this. The dimensions covered by this framework are summarised below (c.f. Figure 2.1: *Balanced Assessment in Mathematics: a Framework for Balance*).

Task type

This attempted to summarise the main purpose of the task, and to justify why someone might be faced with it in the real world.

3 - Assessing problem solving skills: a case study

- Design or Plan
- Evaluate, Optimise, Select
- Model and Estimate or Deduce (from descriptions or images)
- Deduce from Data
- Review and Critique
- Find Relations
- Translate, Interpret & Re-Present Data

Content/Curriculum knowledge

For the *World Class Tests* project, the content was pre-defined as:

- Mathematics
- Science
- Technology

The limited time allowed for assessment and lack of emphasis on curriculum knowledge precluded any fine-grained coverage within the science or technology domains. Since the majority of tasks had some mathematical content, some attempt was also made to cover a spread of mathematical topics (number, shape and space, algebra/formulation, logic etc.)

The “upper limit” on assumed knowledge was taken from the National Curriculum for England and Wales for the level which the candidates were already expected to have attained. Any knowledge above this level had to be introduced by the task itself.

Context type

This broadly described the context in which each task was set:

- Student Life
- Adult Life
- The School Curriculum
- No external context

This needed to be balanced to ensure that the overall test was relevant to the experience of students. Less familiar contexts would tend to make the task more challenging, even if the underlying principles were familiar. For the *World Class Tests* project, which did not focus on numeracy or “functional mathematics”, abstract or fantasy contexts were included.

Practicality

Even tasks set in a familiar context might appear irrelevant or un-engaging to students if the goal or purpose behind the task is abstract or not obvious (for example, almost any pure mathematical number puzzle might be presented as a child performing a magic trick – a

3 - Assessing problem solving skills: a case study

useful technique, but one which could be overused). This was assessed on a 10 point scale ranging from “*immediately useful*” to “*provides insights and methods which may be useful in the future*”.

Openness

Assessment questions commonly have a well defined “correct” solution (often implicit in the style of question, if not explicitly stated). This is atypical of many problems that occur in real life.

Truly open-ended tasks (in which both fully defining the problem and finding a solution form part of the task) are difficult to incorporate in an assessment test, due to time constraints and the need for systematic marking. However, any problem solving task requires an open middle where some non-routine search for solution strategies has to be made.

Tasks may also ask for *multiple solutions* which experience has shown to be challenging for students.

Reasoning length

The ability to construct *substantial chains of reasoning* is a vital aspect of problem solving – yet there is a tendency in mathematics assessment, as exhibited by both GCSE (see Chapter 5) and PIM (see Chapter 4), to break longer tasks into small, prompted, sub-tasks. The *reasoning length* is the estimated time required for the longest prompted sub-task within a question (usually indicated by a numbered question and/or space for an answer) .

Phases

This attempts to characterise the relative demands of each task in terms of five generalised stages of solving a problem:

- Formulating
- Processing
- Interpreting results
- Checking results
- Reporting

Tests were constructed and validated against the above domain specification using an adaptation of the same “balancing sheet” technique developed for *Balanced Assessment*. A sample balancing sheet is shown in Figure 3.1 (c.f. Figure 2.2).

3 - Assessing problem solving skills: a case study

WCT Problem Solving		Age 9 Test 2:1							Balancing Sheet						
© MARS 9July02		Total tasks	Total marks/ mins	Eco-Puzzle	Licence	Shape Factory	Powders	Towel	Pop Stars	Bat Fright	Space Stickers	Making a Shed	Towers	Money	The Race
Task Name		0	107	7	12	10	10	10	5	8	7	10	10	10	8
Marks		0 avge wt>>													
Weight factor															
Strategic aspects		tasks	marks												
Task type															
	Design or plan	2		1	1							1			
	Model, estimate, predict	1					1								
	Select: evaluate and recommend	0													
	Critique and review	0													
	Deduce from data, fit constraints	4				1			1				1		1
	Discover or infer relationships	4			1	1					1			1	
	Translate: interpret & re-present	2						1				1			
	Other	0													
Non-routine-ness: method or inference context		12		1	1	1	1	1	1	1	1	1	1	1	1
		10			1	1		1	1	1	1	1	1	1	1
Open-ness															
	open-middle	11		1	1	1	1	1	1		1	1	1	1	1
	multiple solutions	1											1		
	open-ended	3		1	1					1					
Reasoning length			69	5	6	4	5	4	5	5	6	10	6	5	8
Practicality															
	practical impact			1	1			1	1	1	1	1	1	1	1
	insight					1	1								
Context type															
	student life	5		1					1	1	1				1
	adult life	3			1			1				1			
	curriculum	3				1	1						1		
	no external context	1													1
Phases															
	exploring, experimenting, planning			3	5	5			2	2	2		3	2	2
	manipulating, transforming, making			7	2			6	5	5	3	5	4	4	2
	inferring and formulating theories			3	3	5	5							4	2
	reviewing, checking, testing			2					3						
	reporting, presenting and explaining							4		3	5	5	3		4
Subject focus															
	Mathematics	7			1	1		1	1			1	1		1
	Science	2		1			1								
	Technology	4			1	1				1		1			
Content - more details															
	Number, Quantity, Measurement	5						1			1	1	1	1	
	Algebra and Function	4			1	1					1			1	
	Geometry, space and shape	1										1			
	Data, statistics and probability	0													
	Other mathematics	4			1	1			1						1
Science															
	Experiments, Evidence, Hypothesis	2		1			1								
	Physics	0													
	Biology	2		1			1								
	Chemistry	0													
	Earth Sciences	0													
	Social sciences	0													
Technology															
	Design knowledge	3				1					1		1		
	Properties of Materials	1									1				
	ICT	2			1	1									

Figure 3.1: A "Balancing sheet" used during the development of World Class Tests

3.3: The role of the computer

Although the original brief called for an entirely computer-based assessment, the consensus of the designers was that the “state of the art” of computer-based testing and automatic marking would require highly structured questions with constrained response formats, precluding the type of open-ended, unstructured tasks which formed an essential component of the *Balanced Assessment* philosophy. The arguments for this were similar to those presented in section 2.5. QCA accepted this and it was therefore decided that each test should consist of two parts – one using pencil-and-paper and another delivered by computer.

In addition to the pencil-and-paper-only tests, the computer-based tests would also be accompanied by a paper workbook. For the mathematics tests, these were used purely to provide space for rough working. In the case of problem solving, however, some on-screen questions would instruct the students to write the response in their workbook. This was seen as the only way that students could respond to questions which required a description (possibly including mathematical notation) or demonstrate that they could, autonomously, choose to represent data as a chart or table without being given an on-screen form which defined the format for them.

Although probably untenable in the long term for a “computer based” assessment, this did provide a valuable interim solution as task styles developed. It was also the only way that tasks could be trialled in the early stages of the project, before the data collection infrastructure was in place. Towards the end of the project, as experience was gained by the designers, the dependence on the answer books was waning. Had task development continued, the answer books would probably have been dropped or, as with the mathematics tests, relegated to “rough work” which would not be marked.

The availability of the written paper meant that the computer tests did not have to waste effort replicating tasks that were known to work well on paper, and could concentrate on ideas that exploited the computer to the full. The answer booklet for the computer test meant that the computer could be used to present contexts and information in an interactive format without sacrificing the ability to ask less structured, investigative questions.

Qualities that made a task particularly suitable for use in the computer-based component included:

- The use of animation or interactive graphics to present concepts and information that would be hard to communicate, in simple language, on paper
- The provision of a substantial data set, for students to explore with searching or graphing tools

3 - Assessing problem solving skills: a case study

- Use of simulated science experiments, games and other “microworlds” - allowing question types that would be impossible on paper
- Other types of question that were more suited to computer than paper – for example, questions that naturally suggested a “drag and drop” interface

The main constraint was that the test was to be assembled from self-contained, 5 to 15-minute tasks. Although such tasks are long compared to those typically found on current mathematics tests, it is quite short for the sort of open-ended investigations suggested by the criteria above. As well as the direct limitation on task length, this meant that any on-screen “tools” which the pupil was expected to use within a task had to be extremely simple and intuitive to operate, otherwise valuable assessment time would be wasted on on-screen tutorials and practice before each task.

As the tests were to be scored and graded conventionally, each task also required a systematic, summative marking scheme (rather than the sort of holistic judgement-based scheme tried with the early *Balanced Assessment* tasks – see Section 2.3) so even without the constraints of capturing the answer on computer there needed to be a definite “outcome” against which performance could be reliably assessed.

The other constraint was that tasks had to be produced in significant quantities (over the course of the project, 110 computer based tasks were developed, each representing 5-15 minutes of assessment and usually involving some sort of interactive animation or simulation). This limited the amount of software development effort that could be devoted to an individual task.

3.4: Illustrative examples of tasks

Working versions of these are available on the Appendices CD.

Simulated experiments and “microworlds”

One of the challenges for the problem solving strand was to cover the field of “problem solving in science” without depending on advanced curriculum knowledge – a particular problem at age 9. The computer allowed the presentation of simulated science experiments – in a simplified but defensible form – that embodied all the required knowledge and left students to investigate, draw inferences and justify their conclusions. Figure 3.2 shows one example, which allowed 9-year-olds to successfully engage with the beginnings of Archimedes' principle, eliciting insightful responses such as:

“All the vegetables and fruits that sinks overflow less than they way. All the food that float overflow how much they way”

The screenshot shows a digital interface for a science experiment titled "Floaters". At the top right, there are two buttons labeled "Page 1" and "Page 2". The main instructions are:

1. Place each food on the scales. What is its mass?
Now put it in the bowl of water. How much water overflows?
Does the food sink or float?
 Write down your results in the table in the workbook.
2. Write about any patterns you can see in your results.

The visual elements include a digital scale showing a mass of 80g with a carrot on it. Next to it is a blue bowl of water with an orange floating in it, and a measuring cylinder showing 160 cm³ of water. Below these are several other food items: a green apple, a potato, a red tomato, a bunch of green grapes, a green pear, and a yellow banana.

Figure 3.2: Floaters - a simulated science experiment

3 - Assessing problem solving skills: a case study

The task *Sunflower*⁶ required students to find the optimum combination of nutrients to grow a giant sunflower. Here the “science” content was imaginary (although plausible) and the underlying task was to perform a systematic search for a maximum, while showing the ability to work with decimal fractions to 2 places.

Table 3.1 shows a “heuristic inference” mark scheme for this task, which allows fully automatic marking based purely on the amounts of “plant food” chosen by the pupil for their best attempt.

Figure 3.3: *Sunflower* – systematic search for an optimum

Amount of A and B for best height achieved	Inference	Score
$11 \leq A \leq 12$	<input type="checkbox"/> Has held B constant while varying A <input type="checkbox"/> Has tried 0 or <1 for B <input type="checkbox"/> Has searched for maximum using integers	+1
$11.0 < A < 12.0$	<input type="checkbox"/> Has used decimal fractions.	+1
$0 < B < 1$	<input type="checkbox"/> Has used decimal fractions less than 1	+1
$0.3 \leq B \leq 0.4$	<input type="checkbox"/> Shows some sort of systematic search for B <input type="checkbox"/> Has held A constant	+1
$0.30 < B < 0.40$	<input type="checkbox"/> Has gone to 2 decimal places.	+1
$A=11.5, B=0.36$	<input type="checkbox"/> Full marks!	+1

Table 3.1: Scoring *Sunflower* by inference

⁶ The idea for *Sunflower*, and several other tasks used in this project, was based on software produced at the Shell Centre for Mathematical Education in the 1980s (Phillips, 1985)

Mathematical games

The tests were not limited to “real life” problems and included several “Number games” such as the example in Figure 3.4. This type of game (a variant of “Nim”) has the advantage that there is an easily accessible optimum strategy. However, it was soon clear that leaping directly to formulating the strategy was beyond the ability of most students, so these tasks typically fell into the pattern:

- Here are the rules – play a few games against the computer.
- Here is the last stage in a sample game – identify the winning move
- Here is another sample game – identify the two moves needed to win
- Now describe the strategy for always winning the game.

Game of 20

Page 1 Page 2 Page 3 Page 4 Page 5

This is a simple game for two players.

- Players take it in turns to cover up a number on the board with a counter.
- The covered numbers are added together.
- The first player to make this total exactly equal to 20 wins the game.

2. Imagine you are playing the game. Six counters have already been placed.

Where will you place your next counter to be sure of winning the game?

Drag the counter on to the board.

Figure 3.4: Game of 20

3 - Assessing problem solving skills: a case study

In *Factor game* (Figure 3.5) the computer played a key role in explaining the rules of the game⁷ using an animated sequence. The student's ability to formulate a strategy was put to the test by challenging them to beat the computer by the greatest margin possible. As a follow up, their understanding of the strategy was probed by asking them to imagine a variant of the game with 50 cards instead of 10 and to suggest the best opening moves.

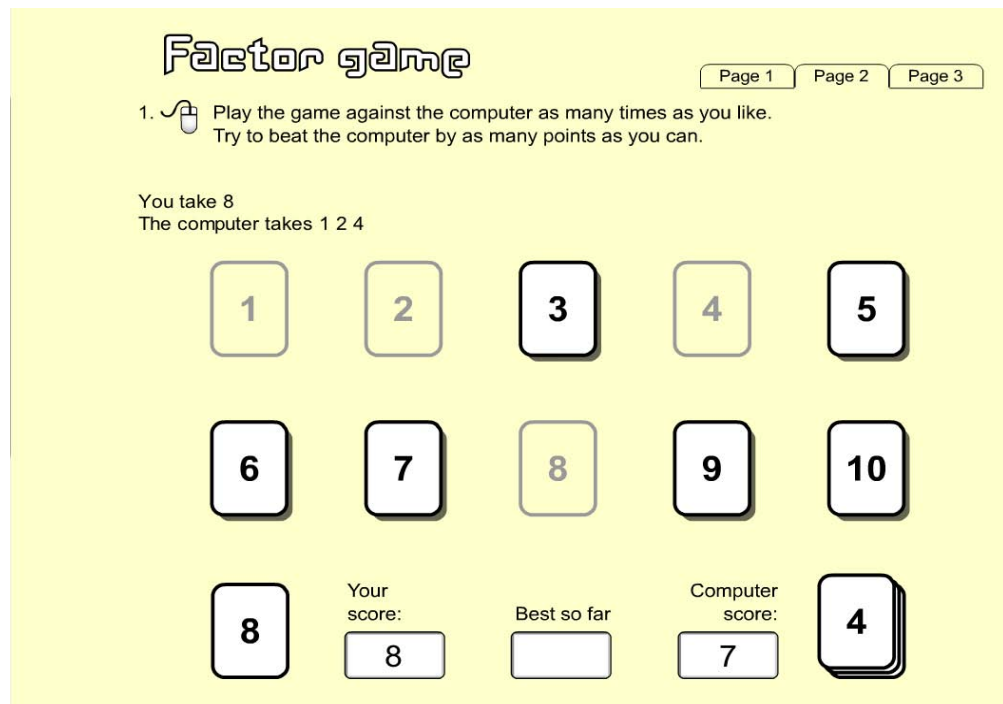


Figure 3.5: Factor Game – human vs. computer

Exploring rich data sets

One advantage of computer-based tasks is that the student can be offered a substantial database, rather than the dozen-or-so cases feasible in a paper test. This allows assessment of the important processes of choosing appropriate data, representing, summarising and interpreting it. *Queasy* (Figure 3.6) requires students to solve a food-poisoning mystery by making suitable queries to a simulated database while *Water fleas* (Figure 3.7) allows a large set of experimental results with several variables to be viewed as bar charts and asks whether these results support or refute a series of hypotheses.

⁷ The rules are: The player picks up a numbered card. The computer then takes all the cards which are factors of the player's number. The player then picks another number, but this must have at least one factor left on the table. Play continues until none of the cards left have factors showing, at which point the computer takes all the remaining cards. The winner is the person who has picked up cards with the highest total face value. The sequence clarified these rules by working step-by-step through an example game.

3 - Assessing problem solving skills: a case study

Queasy Page 1 Page 2 Page 3 Page 4

Database search

Count { all the children
 the children who are ill
 the children who are not ill } (Drag food here) **Result:**

who ate: **Go:**

apples	bananas	cereal	burgers	chicken
chocolate	chips		curry	ham
ice cream	peanuts	peas	sausages	noodles

- Use the database search to find out how many of the children who are ill ate pie. The answer should be 2.
- Now use it to answer these questions:
 - How many children felt ill altogether?
 - How many children ate curry?
 - How many children who ate peanuts also felt ill?

Figure 3.6: Queasy - exploring a database

Waterfleas Page 1 Page 2 Page 3

Craig used his results to draw bar charts.

Temperature: Cold Normal Warm
Pollution: None Some Lots

Number of water fleas swimming

Time (hours)	Number of water fleas swimming
0	100
1	95
2	90
3	85
4	80
5	75
6	70
7	65
8	60
9	55
10	50
11	45
12	40
13	35
14	30
15	25
16	20
17	15
18	10
19	5
20	5

Click on the buttons to see the charts for different levels of temperature and pollution.

If there is no pollution, temperature doesn't have any effect on the water fleas. Craig

- In your workbook, say how Craig's bar charts show that he is wrong.

Figure 3.7: Water fleas – scientific argument

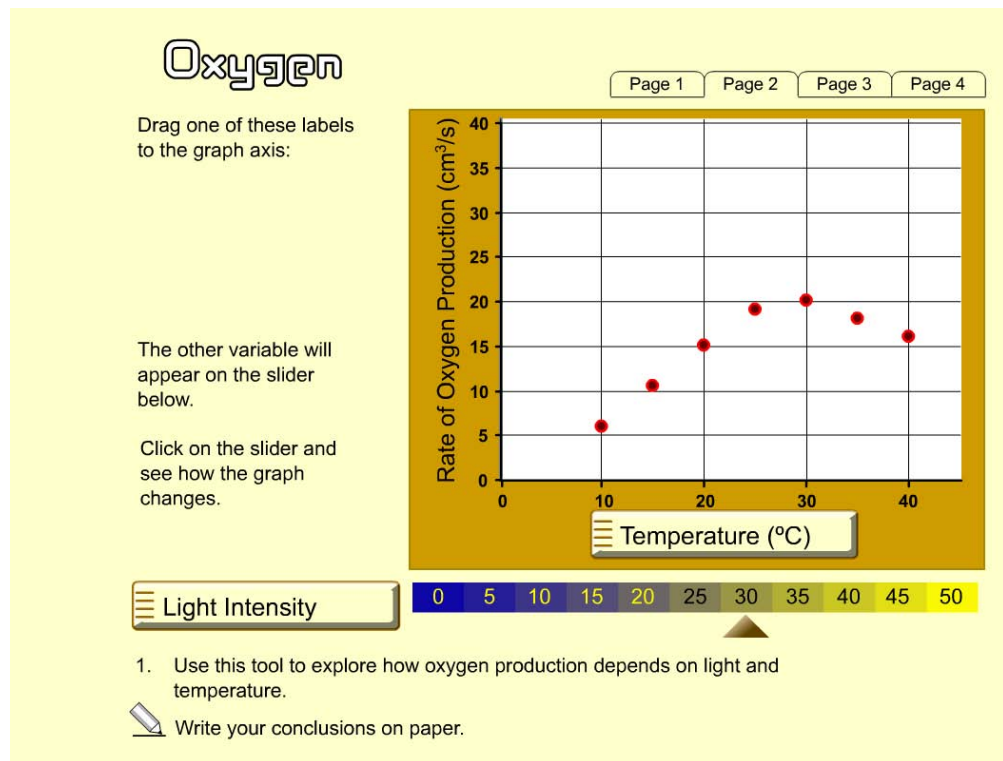


Figure 3.8: Oxygen - exploring multivariate data

Use of the workbooks

As can be seen from the example screens, where questions required a substantial written answer, students were directed to answer in the paper workbook. While this could have been replaced by a type-in text box, this would have placed a constraint on the type and format of answers possible. For example, the task *Bean Lab* (Figure 3.9) reproduced a common classroom science experiment (with a zero-gravity twist not so common in the classroom). The subsequent examples show the diversity of student responses to the first part of this question.

Figure 3.10 is a purely written answer, but the formatting provides valuable evidence of a systematic approach. Producing this “hanging indent” format in a basic, type-in-text field on a computer would have been, at best, tedious and distracting. The test system would have to provide word processing facilities and the students would need to know how to use them.

Figure 3.11 shows a tabulated response, also providing clear evidence of systematic work and good choice of representation. Again, this would have been complicated for the candidate to replicate on computer, and providing a pro-forma table to fill in would have distorted the question. (The work books used “squared paper” throughout to avoid giving any clue that a table or diagram was expected for a particular question).

3 - Assessing problem solving skills: a case study

Figure 3.12 uses sketches which would obviously have been difficult to capture on a computer.

It can be seen from these examples that each student went on to produce a purely verbal answer to the second part of the question, where they are asked to draw a hypothesis from the data. This could have been typed in as plain text, so it might have been possible to discard the answers for part 1 as “rough work” and infer from part 2 whether systematic records had been kept. However, there are two disadvantages with that approach. Firstly, part 1 is an opportunity for less able students to gain some credit for methodical work, even if they are unable to articulate a hypothesis. Secondly, students might have taken less care with this part of the task if they had known that it would not be collected and marked (to properly investigate the significance of this effect would be an interesting future study).

Bean Lab Page 1 Page 2

Here are some experiments to see whether light or gravity affect the way beans grow. One beaker is on Earth. The other is on a space station where there is no gravity.

1. Choose one of the beakers. Set the lights to "Top", "Bottom" or "Off". Drag a bean into the beaker and watch it grow. Try several different experiments.
2. Write down all of your results in your workbook.
3. Describe carefully how light and gravity affect the way the beans grow. Say clearly how your results show this.

Earth **Space station**

Light: Off Top Bottom Light: Off Top Bottom

Figure 3.9: Bean Lab – scientific argument

3 - Assessing problem solving skills: a case study

1. Try several different experiments.

Write down all your results.

Earth - off (Light) - When there is no light, the shoot goes up and the root goes down

- top (Light) - When the light is turned on top, the shoot grows directly to the light (top), the root grows down

- bottom (Light) - When the light is turned on the bottom, the shoot grows directly to the light (bottom), the root grows down.

∴ The shoot always grow toward the light. Gravity makes the roots grow down

Space Station - off (Light) - Where there is no light and no gravity, the shoot and roots grow almost anywhere

- top (Light) - When there is light and zero gravity, the shoot grows toward the light and makes the root grow down

- bottom (Light) - When there is light turned on the bottom, the shoot grows to the light and because there's no gravity the roots grow up

∴ The shoot grows toward the light in the space station, and makes the roots grow in the opposite direction.

2. Describe carefully how light and gravity affect the way the beans grow.

Say clearly how your results show this.

At the earth, the roots always grow down because of the pull of gravity. The shoots grow toward the light.

At the space station, the roots grow in an opposite direction to the light. (there's no gravity) The shoots grow toward the light

Figure 3.10: Bean Lab - written answer

3 - Assessing problem solving skills: a case study

1. Try several different experiments.

Write down all your results.

	Off	Top	Bottom
Earth	The plant grows regularly with the shoot goes up and the root goes down. Gravity makes it grow this way. And there is no light pulling it down.	The plant grows regularly as the shoot goes up and roots go down. Gravity helps do this along with the shoot going towards the light.	The plant grows irregularly as the shoot and the root grow down. Gravity pulls it down as gravity pulls it down and it grows towards the light.
Space Station	It grows irregularly as there is no gravity pulling it anyway and no light for it grow to.	It grows regularly as there is no gravity to pull so it grows towards the light.	The plant grows upside down as the gravity does not affect it and the shoot grows to the light.

1.1 Partial credit : 5 experiments recorded

Directions not specified

2. Describe carefully how light and gravity affect the way the beans grow.

Say clearly how your results show this.

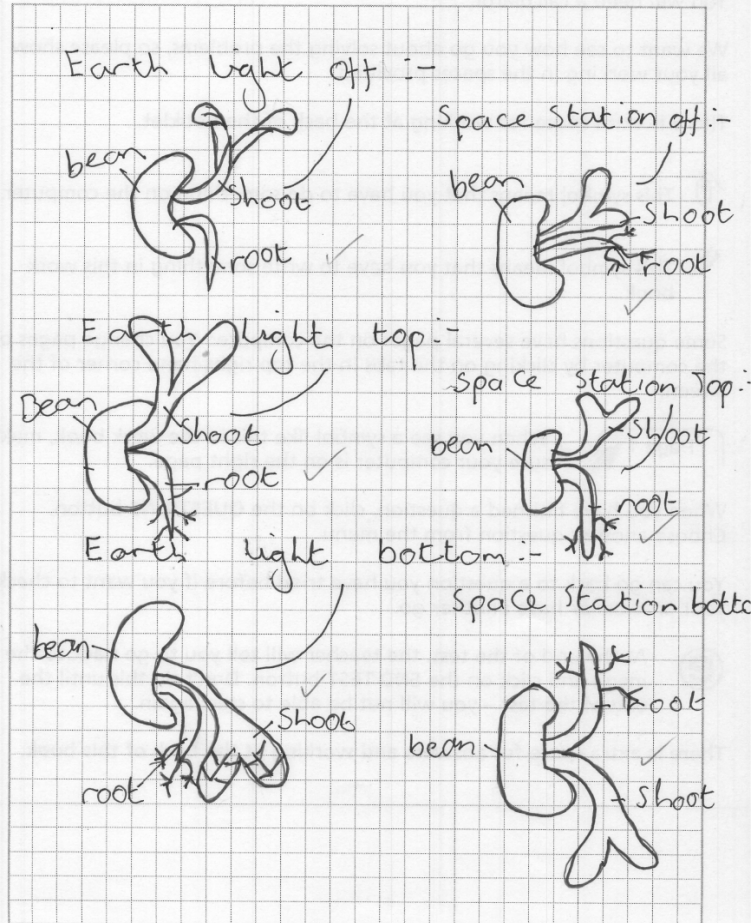
Gravity will make the beans grow properly in the absence of light. If the light is at the top ~~it will be no~~ gravity will have no effect. If the light is at the bottom gravity makes the root point down as well as the shoot. Conditions not specified

Figure 3.11: Bean Lab – tabulated answer

3 - Assessing problem solving skills: a case study

1. Try several different experiments.

Write down all your results.



2. Describe carefully how light and gravity affect the way the beans grow.

Say clearly how your results show this.

because when the light shines from the top ^{off} the bean it grows right but when the light is off it grows wrong.

when the light shines from the bottom, the shoot goes towards the light.

Figure 3.12: Bean Lab - diagrammatic answer

3.5: The development process

Initial design

The design philosophy was that design should start with valid and engaging tasks that would allow candidates to show “what they know, understand and can do” (Cockroft, 1982). Small-scale school trials of the tasks took place at an early stage to ensure that students could engage with the task and demonstrate progress. The marking schemes were developed continually throughout the trials to ensure that they reflected the type and variety of valid responses produced by real students, not simply the designer's anticipated solution, and could be applied reliably by the markers. The balancing instruments described above were then used to assemble a test that adequately sampled the assessment domain.

This approach differs from most test development, which is typically centred on detailed, but abstract, specifications of the curriculum areas to be covered, around which the tasks are constructed. This is straightforward, but can lead to the sort of fragmentation and contrived contexts observed at GCSE (Chapter 5).

The above context-led technique would be impractical if applied universally, so some tasks were inevitably written to address gaps in coverage or balance as the test was assembled.

Ideas for computer-based tasks arose in various ways. They were developed in brainstorming sessions; invented by individual designers and other contributors; adapted from past projects or “appropriated” from tasks under development for the paper test. It was then up to the computer task designer to develop the ideas into a workable specification.

At this point, one of the challenges of computer-based task development became apparent: traditional paper-based tasks at this stage of development would have been drafted, with clip-art graphics and rough diagrams where needed, ready for further discussion and refinement, initial approval by the clients and informal trials. For computer-based tasks, though, all that was available was sketches of the screen layout, the wording of the question and technical notes on how any interaction or animation would work. Tasks in this state could not be trialled in school. Even soliciting feedback from colleagues and clients proved difficult when the task had significant graphical, animated or interactive elements which any reviewer would have to visualise based on the outline specifications.

Specification, commissioning and implementation

Programming of tasks was conducted by a third party, so the next step was to specify each task in detail for the programmers.

The specification had to cover such aspects as:

3 - Assessing problem solving skills: a case study

- Details of the artwork required – this needed tight specification due to the danger of introducing additional clues or distractions (an issue illustrated in section 4.5)
- Details and timings of any animation required
- Careful specification of all the interactions involved, how on-screen objects should respond to various inputs, covering:
 - Suggested algorithms where objects have to move according to mathematical rules, or where the computer must play or referee a game
 - The range of possible inputs for type-in fields (e.g. text, integers, decimals, including the number of decimal places). Should the candidate be warned of/prevented from entering invalid values?
 - Rules for drag-and-drop elements – where on the screen do objects start? How many are available? How they can be removed? Should they automatically align to a grid?
- Details of what data should be captured and stored so that it could be marked
- Details of how the task should be paginated and whether some elements should appear on all pages. This could be crucial, because of the limited amount of information that can be presented on each screen
- Eventually, specifications for the algorithms needed to mark responses automatically, although this stage came after the initial implementation, once a manual markscheme had been designed

In the context of a 10-minute assessment task, where the candidate must be able to rapidly grasp the concept without additional help, the considerations above can be critical and are hard to separate from the educational design. For example, the task designer might design, on paper, a “cloze” question comprising a text with missing words (or numbers) and a list of possible words to go in the gaps. The pupil would copy the correct words into the gaps. A programmer might decide to implement this by displaying the candidate words on icons which the pupil could drag and drop onto the gaps in the text. This is **not** necessarily the same problem, since the new design means that you can only use each word once – a constraint which is not present in the paper version. Even if the correct answer only uses each word once, it is possible that a common mistake involves re-use of a word, so denying the pupil that option could affect the nature of the task.

From the point of view of a software designer aiming to produce robust and easy to operate software, checking the validity of data and dealing gracefully with any unexpected inputs is an important consideration. Adding constraints and checks to the user interface which restrict

3 - Assessing problem solving skills: a case study

the domain of possible responses with which the software must cope is therefore an attractive technique⁸ which might make the task simpler to mark by preventing ambiguous inputs, but could also make the task easier by alerting the candidate when they entered a wrong answer. The educational designer must be involved in deciding how such constraints might alter the question. So, in the above “cloze” example, the designer must remember to specify whether there should be more than one of each icon, something which they might not consider in a paper-based task.

Typically, the first implementation of a task by the programmer had serious faults and one or two rounds of improvement requests were required to arrive at a version ready for trials. This was not simply due to mistakes by the programmer, but often because the designer wished to refine details having seen the first working version. Good communication between the educational designers, graphics designers and programmers was essential here, and the strictly partitioned approach imposed by the *World Class Tests* project structure, where (for instance) change requests sometimes had to be submitted in writing without face-to-face contact with the programmer, was not ideal.

As the project progressed, it was often found to be simpler for the designer to produce partial working prototypes which implemented the critical interactive aspects and included draft graphics and animations, which could be fine-tuned before submission.

In the initial stages, the delivery “shell” which allowed the candidate to log on and navigate through the questions was also under development, as was a “library” of standard buttons, input boxes and other controls. An example of the sort of issue that arose here was whether it should be possible for a candidate to return to a previous question to review, and possibly modify, their responses. This is something that would be taken for granted on paper, but which is only possible on computer if it has been specifically provided for in the test delivery software.

Trial and refinement

Each task was scheduled to go through at least three rounds of trials:

- Informal, closely observed trials with a small number of students to ensure that they could engage with the task and to identify any bugs or shortcomings in either the task content or its technical implementation. These trials were often conducted with students working in pairs, with no attempt made to present balanced tests or gather psychometric data. Working in pairs encouraged students to discuss their thinking

8 From a pure user interface design perspective, a “good” on-line test would, of course, have all the correct answers filled in automatically as a convenience to the user, an approach which would undoubtedly raise performance, if not standards.

3 - Assessing problem solving skills: a case study

(and, sometimes, express their frustrations) without the observer having to interrupt with questions.

- “Formal” trials, with around 50 students taking each task, to establish that the tasks were performing well in an assessment environment and producing an adequate spread of results. These trials remained focussed on individual tasks. The resulting student work was used to refine the mark schemes and to inform the assembly of the tasks into balanced tests.
- “Pre-test” trials of complete, balanced tests – aiming for around 200 students per test – intended to provide statistical data for calibrating the tests.

A major tension was that, for the first two rounds of trial to be worthwhile, it had to be possible to rapidly revise and re-trial a task. There was a conflict between the need to schedule school visits for informal trials in advance and the requirement to commission any revisions from the developers. A flaw in a task might become obvious the first time a child tried to complete it, but whereas a paper task could be redrafted overnight, it was often impossible to revise the software in time for the next scheduled visit. Combined with the delays in task commissioning noted above, and the problems with getting infrastructure in place for trials (discussed below) this meant that it was often impossible to put computer tasks through the full, three-stage, iterative trial and refinement cycle, and many tasks skipped the “formal trials” step.

Some design challenges

Finding the task in the context

The desire for rich and interesting contexts has to be balanced with the constraints of the assessment. Many appealing subjects emerged from brainstorming sessions – such as Muybridge's famous pictures of galloping horses, or analysis and comparison of demographic data from many countries – but identifying a self-contained, 5-15 minute task set in that context proved difficult.

One of the hardest decisions for a lead designer was when (and how) to diplomatically reject a contributed idea, into which a lot of research and effort had already been put and which would make a wonderful extended investigation, on the grounds that no well-defined, scoreable task had been identified.

Eliminating trial and error

When designing interactive test items based around a microworld or simulation, a key challenge is finding questions which genuinely probe the students' understanding of the

3 - Assessing problem solving skills: a case study

situation and which can not be answered with a simplistic “trial and improvement” approach in which the student uses the simulation to check possible answers.

Tactics used to eliminate or reduce trial and improvement include:

- **Written explanation** – ask students to describe their strategy/justify their findings, or to support/refute some suggested hypotheses
- **Simple challenge** – ask students to “beat the computer” and rely on the time constraints of the test to discourage brute force/trial and error solutions
- **Logging and analysis** – record every interaction between the student and computer and then try to analyse this data to spot promising patterns and sequences. This requires complex coding and could be fragile: a few random interactions not indicative of the students' thought processes could disguise a valid response. Generally, a large corpus of trial data would be needed to validate such an approach
- **Heuristic inference** – Table 3.1 shows a possible scheme for marking the *Sunflower* task (Figure 3.3) which infers the sophistication of reasoning and strategy shown by the pupil based solely on their best result, without recourse to their written work. Likewise, with *Factor Game* (Figure 3.5) the final score was taken to be indicative of the level of understanding: most students could beat the computer eventually; a “high score” of 30 suggested that the student grasped the idea of factors and multiples; 35 implied they had made some progress towards a strategy for improving their score while the optimum score of 40 was unlikely to be achieved without a well developed strategy. This has the advantage of being easy to program and fairly easy to justify – but the approach does not lend itself to all tasks
- **Extension problems** – after exploring an interactive scenario, such as a computer game, the student is asked to demonstrate their understanding by making inferences or predictions about an extended or generalised variant, with no simulation available. This technique was also used in *Factor Game*, where the final challenge is to suggest the optimum opening moves in a game with 50 cards instead of 10 . In other cases, an arbitrary limit was set on the range of inputs accepted by the simulation and the final question lay outside that domain.

3.6: Technical and Logistical Challenges

Technical issues

The project started before widespread access to broadband internet connections could be taken for granted. Consequently, most of the tests were delivered on CD and had to be

installed on individual computers. The data then had to be extracted from the individual computers and returned by email or mailed on floppy disc.

This proved to be a major challenge – especially in schools with networked systems that prevented individual machines from writing to their local hard drives. Although this potentially meant that administration and data collection could be centralised, the diversity of networking systems and lack of technical support made installation complicated. Even on stand-alone systems there was a high incidence of lost data when teachers were asked to manually copy and return data. The agency performing the programming and delivery software design was also somewhat naïve about the level of technical proficiency that could be expected from teachers (such as their ability to copy files by dragging and dropping rather than opening them in a word processor and re-saving).

Whatever the problems with internet delivery of assessment (see Section 6.11) the possibility of “zero-install⁹” delivery and automatic return of data is attractive in the light of the experiences with *World Class Tests*.

Project management issues

The early years of the project were somewhat fraught, and there may be some lessons to be learned for future projects. Some of the issues included:

- **Structure of the project** – the organisation, as conceived, was heavily compartmentalised – with two groups contracted to work on the educational design, a third contractor handling the software development and a fourth (appointed later) responsible for “delivering” the tests. This seemed to be founded in a publishing metaphor: manuscript -> editor/designer -> publisher/distributor; which assumed that the hand-over between each stage was routine and well understood. Initially, this led to designers being unaware of the constraints of the delivery system and programmers not understanding the aspirations of the designers.
- **Task specification and approval** – as discussed above, when tasks involve substantial interactive elements, programmers must be supplied with more than the question text and a sketch of the artwork. The workload of specifying the tasks, testing implementations and specifying revisions had been underestimated, and largely fell on one or two people. This delayed the commissioning of new tasks from the programmers – who were expecting a steady flow of routine work.

⁹ Applications that run without requiring custom software to be installed – usually using a standard web browser or (by a less strict interpretation) ubiquitous, general-purpose plug-ins such as Flash or Java.

3 - Assessing problem solving skills: a case study

- **Prototyping** – in a non-routine project such as this, it is hugely ambitious to expect to go directly from paper specification to final implementation. Ways of prototyping partly-working examples to try out and rapidly refine – or possibly reject – ideas need to be considered.
- **Technical oversight** – the project had several stages of internal and external review to ensure the educational validity of the materials. There was, initially, no corresponding oversight of the technical issues or agreement between the designers and programmers as to what the constraints or expectations of the system were. An internal committee was eventually set up, but its source of authority was unclear.
- **Timing** – although the overall timescale – two years until the first live sittings - was appropriate, the contract mandated a large scale trial just a few months after the effective start of the project. This would not have been unreasonable for paper based tests which could easily be piloted in draft form, but delivery of computer tasks required substantial infrastructure development as well as programming of the actual tasks, and the attempt to meet this requirement largely failed. Multiple rounds of trial, feedback, revision and calibration are critical to developing a robust and valid test but, in a computer-based project, need to be reconciled with the fact that a substantial amount of programming needs to be completed before any materials can be trialled.
- **Short-term contracts & rights** – this affected the programming side in particular – with no ongoing commitment to continue the contract after the initial two years and all IP rights assigned to the client, there was little commercial incentive to invest time in building a solid IT infrastructure which might then have been taken over by the lowest tenderer at the end of the contract.

3.7: Outcome of the project

The project produced a bank of 5 complete tests at each of ages 9 and 13, which have been successfully administered, marked, moderated and graded on a commercial scale, setting it apart from “blue sky” eAssessment projects that develop and deeply research a handful of ambitious exemplar tasks.

Students in the target ability range were able to make progress on the tasks, producing a good spread of scores which adequately discriminated between different levels of performance.

Development of new test items was stopped in 2003, but test sittings continue with the existing materials – see www.worldclassarena.org. From that site: “Since the first test session in 2001, over 18,000 students in over 25 different countries worldwide such as Australia,

Hong Kong, New Zealand, Saudi Arabia, Slovenia, the United Arab Emirates, the United Kingdom and the United States have taken the tests.”

In the later stages of the project, it was realised that students who had never encountered these types of problem in the classroom found the tests particularly difficult. Consequently, some of the test development effort was diverted to produce teaching materials based around adapted and extended versions of previous test questions. The approach used was that students would tackle the task individually or in pairs, and then engage in a classroom discussion in which they compared their techniques with other groups, and against specimen solutions provided with the materials. The tasks chosen were, intentionally, quite hard so many pupils would only make progress after sharing techniques.

The classroom materials were published by nferNelson, including 6 modules under the title *Developing Problem Solving*.

More details of the design philosophy of these tests can be found in *Computer-based assessment: a platform for better tests?* (Burkhardt & Pead, 2003).

3.8: Conclusions

In 1.2A we asked “How can eAssessment contribute to the assessment of problem solving skills in mathematics?” The *World Class Tests* project shows that the computer can deliver rich, open tasks involving simulated experiments, “microworlds” puzzles and games, significantly expanding the domain of task types and contexts which can be included in a formal, external assessment.

The project also showed that students could successfully investigate and explore relatively complex relationships when they were presented clearly and interactively on the computer – in one study based on the materials (Ridgway et al., 2006) computer-based tasks such as *Water Fleas* (Figure 3.7) and *Oxygen* (Figure 3.8, p41) involving multivariate data were found to be scarcely more difficult than paper-based tasks based on simpler data sets. The implication of this is that students could realistically be assessed using more complex, realistic and relevant problems on modelling and statistical literacy than is possible by conventional means. This is one way in which online assessment could “improve the range and balance of the assessed curriculum” - the question raised in 1.2C.

The main success of *World Class Tests* was in using the computer to deliver microworld-based tasks in a mixed computer and paper assessment. However, half of the assessment in *World Class Tests* was still in the form of paper-and-pencil tests, in addition to which the problem-solving computer tests relied partly on a paper answer booklet. While the challenges in producing a completely paperless test may have been soluble on a task-by-task basis, the

3 - Assessing problem solving skills: a case study

design and programming load of scaling this to adequately sample the subject domain and deliver 2-4 test sittings a year would have been considerable.

The greatest implication for the “technical and pedagogical processes of computer-based assessment design” (1.2D) is the clear need for two, usually separate, areas of expertise to work together to ensure that the technical aspects of the product reflect the pedagogical principles on which it was based. Task designers accustomed to handing over their paper manuscripts for conventional typesetting and printing need to become involved in key decisions over animation, interactivity and response input methods, while programmers need to learn how their decisions can impact on pedagogical issues and know when to refer a technically-driven change back to the designer . If programmers are to work from detailed specifications then it must be recognised that developing these specifications is a new and significant phase of development not present in a traditional paper-based product cycle.

There are also challenges for design research models which rely on multiple, rapid cycles of trial and refinement: this is straightforward when the “refinement” step means a few changes to a paper document; less so when it entails specification, commissioning and testing of software changes.

4: Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

4.1: Introduction

While projects such as *World Class Tests* set out to produce new assessments tailor-made for computer-based delivery, it is inevitable that other projects will seek to build on past investment and experience by creating computer-based versions of existing paper-based assessments. These may have a long-established reputation or have been validated and calibrated through large-scale trials. Is it safe to assume that the new, computer-based tests will automatically inherit the validity of the original tests? Even if the tests remain valid, is it reasonable to compare scores directly with the paper-based originals, or will it be necessary to completely re-calibrate the tests?

nferNelson¹⁰ produces a series of mathematics assessments under the title *Progress in Mathematics* (PIM). These are also available in a “digital” version using a proprietary online test delivery system. While 20-25% of the questions on the digital test are new, the majority were developed directly from questions from the paper test.

In 2005 nferNelson carried out an “equating study” to compare the performance of the digital and paper versions at ages 6, 7 and 11. The results were generally positive, although it was observed that the means scores on the digital questions were consistently lower than for the paper equivalents.

¹⁰ The tests were developed by the National Foundation for Educational Research and published by nferNelson – now known as “GL Assessment”. The business relationship between these entities is beyond the scope of this thesis.

The author was commissioned in 2006 by nferNelson to investigate the significance of the results of the equating study, conduct observations of pupils using the test, and establish whether the computer versions were “truly equivalent” to paper, to inform future development of digital tests.

The full report to nferNelson (Peard, 2006 - approx. 150pp) including the full task-by-task analysis is available on the accompanying CD-ROM.

4.2: The *Progress in Maths* tests

Progress in Maths (the paper version) is a series of 11 tests, developed and published by nferNelson as a tool to allow teachers to monitor pupils attainment in mathematics at ages 4-14. This study concentrated on the tests for ages 6,7 and 11.

Each test consists of a Teachers's guide, (e.g. Clausen-May, Vappula, & Ruddock, 2004a) and a pupil's work booklet (Clausen-May, Vappula, & Ruddock, 2004b). The teacher's guide contains instructions for administering, scoring and analysing the tests, although the publisher also offers scoring and analysis services and supporting software.

At age 6, the test is intended to take 35 minutes and contains 23 short questions, rising to around 29 rather longer questions in an hour at age 14. The tests can optionally be split over two sessions: at the younger ages, the split can occur at any point near the middle of the test, but at age 10 and over, the test is divided into a “calculators allowed” and “no calculators” session.

In the case of tests for ages 8 and below, the questions are administered orally by the teacher (from a provided script), in an effort to remove any dependence on reading ability. The pupil's answer books at these ages contain the images, diagrams and numbers needed to answer the question along with a minimum of text: this means that most of the questions could not be answered without the additional information read out by the teacher.

The questions on each test had been selected and refined based on trials of candidate questions with approximately 200 pupils, plus questionnaires from teachers. Following that, the finalised tests were standardised nationally (across the UK) in 2004, using samples of around 2000-2500 pupils or more at each age. The teacher's guides contain the instructions and data needed to convert raw test scores into age-corrected standardised scores (since the tests will usually be used with a whole school year, variations in the actual ages of pupils are significant, especially in younger age groups) and to predict pupils' ranking with respect to the national sample.

Progress in Maths – Digital Edition is described by the publisher as “a completely computer-based version of the *Progress in Maths* series”. It is delivered online using the publishers'

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

proprietary assessment engine *Testwise*. Tests are marked automatically and reports for teachers generated automatically.

Most of the digital questions are direct conversions of paper questions (for example, of the 28 questions on the PIM 6 digital test examined, 18 questions were clearly copies of paper questions, and of the remaining 10, 3 showed some similarity in context and mathematical content to a paper question).

Differences between a paper question and its digital copy could include:

- Two-colour, shaded illustrations replaced by full colour
- Layout changes to fit on the screen rather than the printed page
- Some two-part questions split over two screens
- “Write your answer on the answer line” replaced by “Click and type your answer in the box”
- “Put a ring around (the correct answer)” replaced by “Click on (the correct answer)”
- “Put a tick on the chart to show...” replaced by “Move the tick to show...”

Key overall differences between paper and digital tests include:

- At ages 6-8, rather than having the teacher read out the questions, the questions are posed by a recorded voice (so headphones are mandatory if a class is taking the test); the question is read out once when it is first displayed and pupils can click on a “listen again” button to hear it repeated
- Tests must be taken in a single sitting – they can not be split over two days or with a break as suggested for the paper tests (especially for younger pupils)
- The “recommended” timings for the paper tests are, in the digital version, strictly enforced at age 9 and above, with an on-screen clock. At ages 6,7 and 8 there is no on-screen clock and the test allows 5 minutes more than the recommended time before forcing pupils to stop
- Each digital test is preceded by a series of “practice” questions intended to familiarise pupils with the user interface.

To establish equivalence between paper and digital versions, the publisher conducted an “equating study” in which around 100-200 students for each year sat both the paper test and its digital equivalent. Table 4.1 shows the correlation between the digital and paper versions as determined by nferNelson. The publishers observed that “in almost all cases, mean scores for digital items are lower” and also that some questions appeared to have a noticeably larger difference in mean score.

	Correlations	Number of pupils
PIM 6	0.74	181
PIM 7	0.71	160
PIM 8	0.82	97
PIM 9	0.85	232
PIM 10	0.89	179
PIM 11	0.85	238
Total		1087

*Table 4.1: Correlation Coefficients Paper/Digital PIM 6-11
(analysis supplied by nferNelson)*

The aim of the work described here was to:

- Further analyse the data from the equating study to identify questions with significant differences between mean scores on the paper and digital versions and investigate possible causes of these differences
- Review the items in the age 6-8 tests with regard to changes made to paper questions in the transfer to computer, and the design of the new computer-only items
- Observe pupils taking the age 6 and 7 digital tests to identify any causes of difficulty
- Survey the attitudes of teachers and pupils towards the digital tests

4.3: Approach

Analysis of the equating study data

Since the equating study had already been conducted by a third party and the experimental design was pre-determined, the first step was to investigate the data for any effects or bias which could affect the publisher's interpretation of the apparent discrepancies in overall mean scores and correlations. Such effects might include:

- **Overall ability** – might the difference in media be disproportionately affecting lower- (or higher-) ability pupils?
- **School-wide effects** – although the number of pupils is respectable, these are distributed between rather fewer schools. Are the discrepancies consistent across schools, or are they concentrated in individual schools? If so, is there any evidence as to the cause?
- **Order and timing of administration** – in the equating study, each pupil took both digital and paper tests. Since 75-80% of questions are recognisably the same on both

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

tests, it was reasonable to assume that these scores would **not** be independent. How large is this effect, and does it bias the overall results?

- **Individual items** – are some questions showing more significant differences between paper and digital scores than others? If so, are these differences dependent on the other possible effects noted above? Such items may indicate particular issues with the design of digital questions.

The overall approach was to start by exploring the data graphically to identify possible effects which could then be confirmed using statistical significance tests.

In the case of possible differences between items, the publishers had already highlighted a number of items with visibly large digital vs. paper discrepancies ($\geq 10\%$ difference). These “rule-of-thumb” observations were investigated using three techniques:

- **Visually** – using profile graphs of item facility levels. This was also used to investigate school-on-school differences.
- **McNemar's Test** – which can be used to compare two pass/fail measurements on the same sample (unlike chi-squared, which assumes separate samples).
- **Rasch scaling** – uses a probabilistic model to place test items on a difficulty scale by comparing the relative odds of a pupil “passing” each item. When the results of both paper and digital tests are scaled together, equivalent items should receive similar rankings.

Design critique of the questions

The digital tests at ages 6,7 and 8 were critically examined to identify design or implementation issues which might, in the case of converted paper items, lead to differences in performance or, in the case of the new digital items, affect their validity.

The types of issues to be considered included:

- **Intentional changes** – small, apparently deliberate changes in the mathematical content of some questions were noted and the possible impact considered. Also, where new “digital only” questions appeared to be replacements for (but not copies of) paper items, the two questions were compared and contrasted
- **Changes in response mode** – even where the mathematical content of the digital and paper questions was the same, the method by which the pupil provided the response was inevitably different (for example, typing a number instead of writing; clicking on the correct answer rather than drawing a ring around it). Did these constrain the possible responses or introduce clues (either helpful or misleading) as to the answer?

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

- **Presentation issues** – could the introduction of colour, changes to layout or additional illustrations influence the performance of the questions?
- **System design issues** – what are the key differences in the overall testing environment and could they influence the outcome? This included issues such as the use of a recorded voice instead of the teacher reading the questions; differences in timing and pacing and details of the user interface design and conventions
- **Technical faults** – are there any bugs which could be impacting performance?
- **Student responses** – where it was possible to extract a list of all the distinct responses, and their frequency, to a question in the equating study, these were examined for clues for changes in performance. A change in the “most popular wrong answer” between paper and computer, for example, could imply that the computer version had introduced (or removed) a distraction or misconception

This analysis included the “practice questions” taken before the main test, as well as the background screens shown while the general test instructions were being played.

School observations

Medium- and large- scale tests such as the equating study produce valuable psychometric data on the performance, both on the test as a whole and individual question items. However, most of these results are in the form of generalised statistics, such as facility values for individual questions – it is difficult to relate these results to specific aspects of task design without significant inferential leaps. This is especially true when the only evidence generated by most questions is a multiple choice or single-number answer.

Close observation of students actually using the materials might reveal insights into how they interact with the questions which could not be deduced by analysing data after the event. Furthermore, since the existing equating study was expected to yield significant psychometric data, there was no need for the observations to be constrained by the need to capture valid scores, meaning that the subjects can be questioned and prompted in an attempt to elicit the source of any observed issue.

The three key questions to which the observations sought answers were:

- Is there evidence of possible causes for consistent under-performance on digital tests vs. paper-based tests (although the evidence for such an effect from the equating study was weak)?

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

- Is there evidence to explain the significant under-performance (or, in a few-cases, over-performance) on specific questions indicated in the equating study, and does it support the issues raised by the design critique?
- Do the new items designed specifically for the digital test appear to be working as intended?

Throughout June 2006, 70 students from 6 schools were interviewed and observed taking the PIM 6 and 7 tests. Schools were chosen from a list of customers for the traditional versions of the PIM tests, supplied by nferNelson, who also offered an incentive to schools who took part, in the form of free vouchers for their products. The observations were conducted by the author and a “helper” (who had extensive experience of teaching and task design).

For each test, schools were asked to select two "above average", two "average" and two "below average" students - the majority of pairs consisted of a girl and a boy. Schools ranged from a small C of E infant school in a relatively affluent area to a larger junior school near a large housing estate. Since this was a qualitative exercise, no other attempt was made to stratify the sample in terms of ability, race or background - to do so would have required a much larger sample spread across more schools.

Students took the test in pairs, and were encouraged to take turns and discuss the questions with each other – the aim was to get them to naturally “externalise” their thinking and reduce the need for the observer to ask questions. Students were given time to tackle each question uninterrupted, but once they had either completed the question or become stuck, the observers would intervene (preferably by asking probing questions rather than revealing the answer) and try to establish whether any difficulties observed might be attributable to the use of the computer.

Surveys and interviews

Before each pair of students took the test, a short, structured interview was conducted in an effort to collect some background information on their attitudes to mathematics, computers and tests. This was based on a series of fixed questions – the observer ringed one or more of a set of possible answers. The answers from each member of the pair were collected separately. A similar series of questions was asked after the pair had taken the test.

Where possible, teachers were also interviewed to determine their attitude to technology and assessment. One more specific series of questions was introduced part-way through the visits to investigate how teachers approached the administration of the paper test, in order to compare the use of teacher-read questions in the paper test with recorded voices on the computer tests.

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

Some notes were also taken as to the environment and IT provision at the school – such as the location of computers (in classrooms or specialised computer suites) and the type of equipment in use.

These interviews and surveys were mainly conceived as an attitudinal survey for the client, rather than a rigorous research exercise, and were sometimes skipped or curtailed to ensure that pupils had time to take the test. Although they were too small to be the basis of any significant inferences, they did identify some questions for possible future exploration.

Relationship between the approaches

In an ideal study, the statistical analysis of the equating study, the design critique and the classroom observations would have been conducted independently and “blind”, so the findings of each could be tested, without bias, against the others. The reality of this particular study meant that much of this work had to be conducted by the author and, during the classroom observations, an assistant. The timing was largely dictated by the availability of schools for observation so, while it is generally true to say that the bulk of the statistical analysis was done first, followed by the design critique and then the observations, there was some overlap.

Hence, it is not possible to claim these approaches as truly independent. In particular, the list of questions showing significant digital vs. paper discrepancies was well known during the design critique, and the issues flagged by the design critique were well known when the classroom observations were conducted.

Given that caveat, the author attempted to approach each part of the investigation with an open mind. All tasks were subject to both design critique and classroom observation, regardless of the data analysis results.

In the final report the findings of the design critique were discussed side-by-side with the results of classroom observations and the data analysis, noting occasions on which the various avenues of investigation either corroborated or contradicted each other.

4.4: Results of the equating study for PIM6 and PIM7

Is there an ability effect?

Figures 4.1 and 4.2 show scatter graphs comparing individual scores on the paper and digital tests. Note that the lines shown are not “lines of best fit” but are a visual aid representing equal scores on the two tests.

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

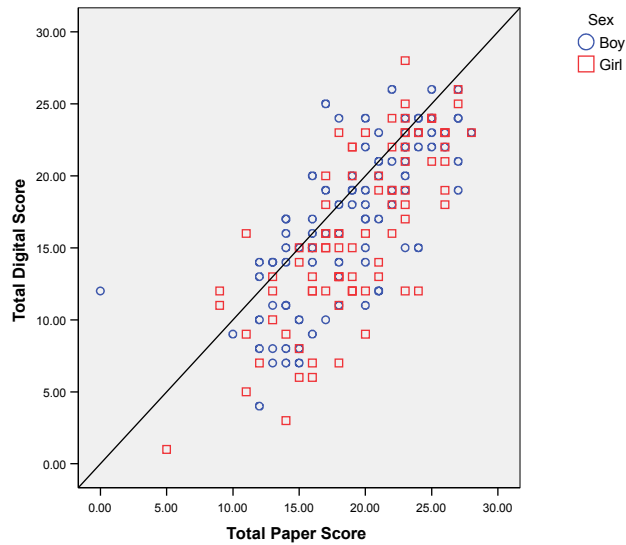


Figure 4.1: PIM 6 Equating study- Paper score vs. Digital score

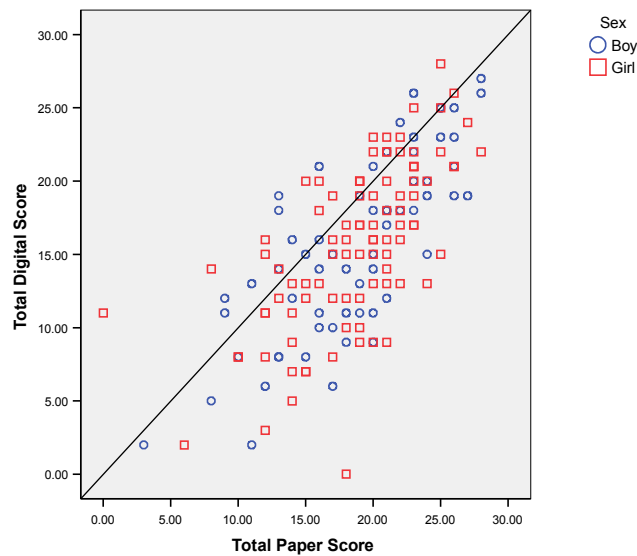


Figure 4.2: PIM 7 Equating study - Paper score vs. Digital score

If the tests were completely equivalent (with just random discrepancies) one would expect the points to be scattered evenly either side of the diagonal line. The fact that there are visibly more points below the line than above it suggests a clear trend of lower scores on the digital test. This graph also suggests that this effect occurs across the whole ability range – regardless of their score on the paper test, the majority of pupils seem to drop a few points on the digital test. Some do considerably worse and only a few do better on digital.

There does not, therefore, appear to be any visually obvious ability effect – statistical analysis might suggest more subtle effects and quantify the effect of digital vs. paper, but since the discovery of the strong effects of different schools and/or order of testing (see below) called into question the validity of whole-sample analysis based on total score, this was not pursued.

Is there a school effect?

When similar scatter plots to Figs. 4.1 & 4.2 above were produced on a school-by-school basis (Pead, 2006, pp. B1-4) it became clear that the effect varied noticeably between schools, with some schools close to “equivalence” and others in which the majority of students scored considerably less on digital. There was a suggestion of two populations – one achieving broadly comparable results on both media and another showing clear discrepancies.

Figures 4.3 & 4.4 show medians and quartiles of total scores by school (by the time these graphs were produced, the role of order-of-testing, discussed later, had become apparent, so they are grouped on that basis rather than the “populations” originally observed). It is quite clear that, while some schools show a clear discrepancy between paper and digital (e.g. school 5213) others showed no effect (e.g. school 2642). It is also notable that, for those two schools, the observation holds for both the PIM6 and PIM7 tests.

This effect would be consistent with either a technical/hardware problem; a non-technical problem (such as a poor environment in the computer facility) or generally poorer student IT skills at the affected schools. The possibility that one or more teachers discussed the questions with students between the tests, or provided extra help when administering the paper test, cannot be eliminated.

It is apparent that a few schools made a disproportionate contribution to the observed differences between paper and digital scores.

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

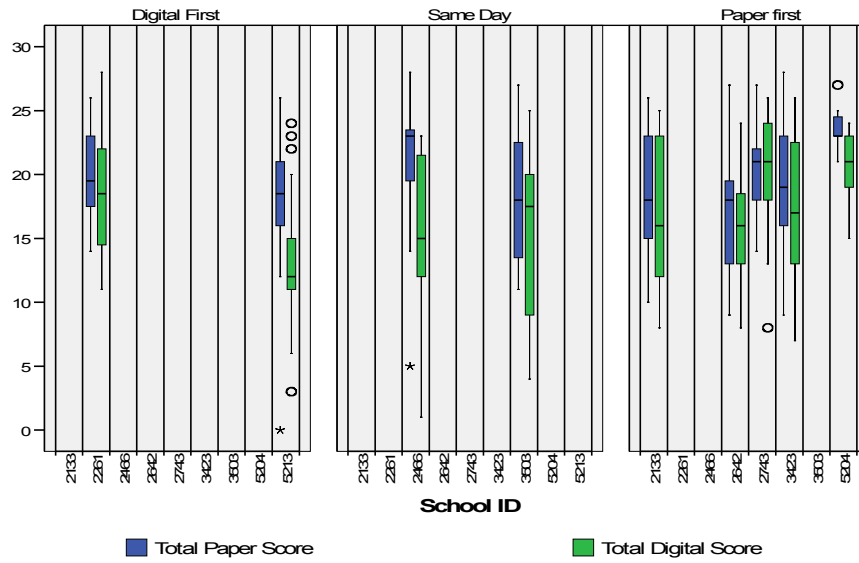


Figure 4.3: PIM 6 Equating study - median and quartiles by school

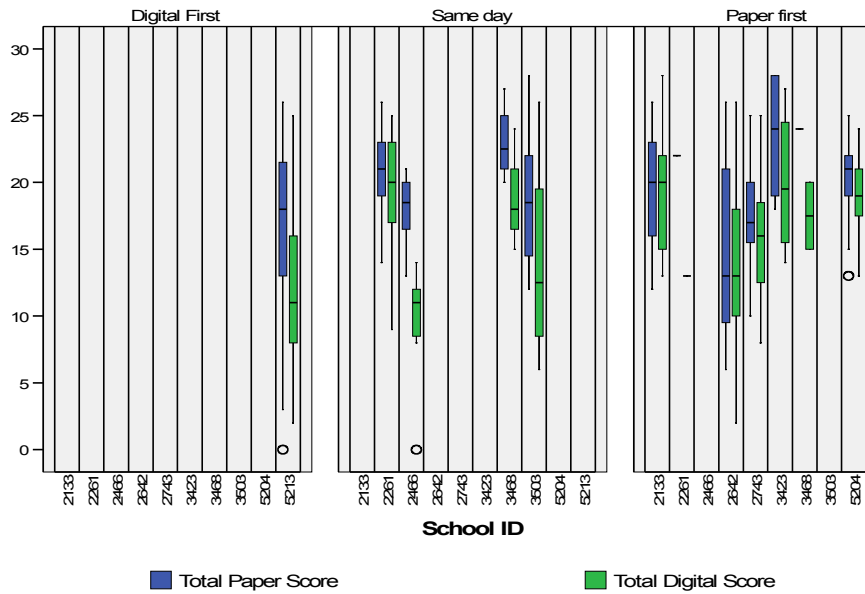


Figure 4.4: PIM 7 Equating study - median and quartiles by school
 NB: schools 2466, 3468 and 3423 each represent <10 cases and should be treated with caution.

Is the order of administration causing an effect?

Once it became apparent that there were two distinct “populations” of schools: those with a large difference between digital and paper scores and those with only a slight discrepancy, more details of the structure of the equating study were sought. It was discovered that there had been no control over which order pupils were given the digital and paper tests, or what the interval was between them. Having established that the raw data contained the date and time on which each pupil sat the test, the possible effect on the order of, and interval between the two tests, was investigated.

Figures 4.5 & 4.6 show the scores discrepancies of individual students plotted against the interval between the two tests. The larger points represent several students at the same point. Students in the lower left quadrant of the graphs took the digital test first *and* scored lower on digital than on paper.

Both of these graphs appear to support the notion that the students who took the digital test **before** or **on the same day** as the paper were more prone to under-performing on digital – and that enough students were in this situation to have a notable effect on the sample.

Figure 4.7 shows scatter plots of digital vs. paper scores, divided into schools who took the digital test before and after the paper version. It is fairly clear that the pupils who took the digital test first tended to perform less well on the digital relative to paper.

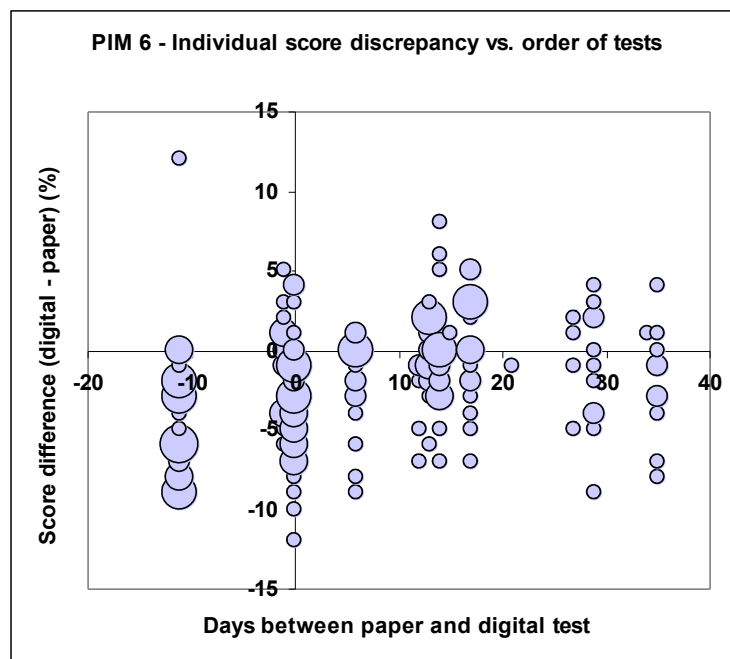


Figure 4.5: PIM 6 Equating study: Effect of order of tests on score difference

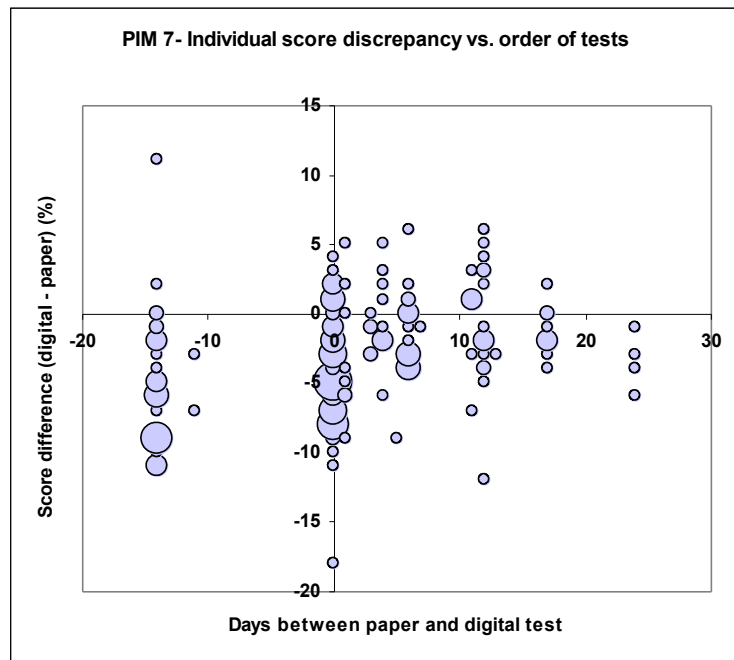


Figure 4.6: PIM 7 Equating study: Effect of order of tests on score difference

An ANOVA test also suggests that the order of testing has an effect on the total mean score on the digital test – but that the effect on the *paper* test is marginal (see Pead, 2006, pp. C-1). This is supported by Figures 4.11 and 4.15, which show little order-of-testing effect on most items – suggesting specific issues with taking the digital test first rather than a general test/retest effect.

Figures 4.3 & 4.4 show box/whisker plots grouped into digital first/same day/paper first which further illustrate this.

However, these figures also show that the “digital first” pupils were all from the same school (age 7) or two schools (age 6). Also, school 2261 took the PIM6 digital test first (they took the paper test the next day) but shows no significant discrepancies. So these results could also be explained by a school-wide effect such as technical problems or poor IT skills teaching. More data – with a more even coverage of test order and interval - would be required to be confirm or eliminate the possibility of a test/retest effect.

The validity of the data from those schools who took both tests on the same day is highly questionable – no information on the order in which those tests were taken was available (it is possible that half a class took the digital test while the other half took the paper version, then swapped). Students would certainly recognise the questions common to both tests – but whether this would be an advantage or a source of complacency and careless mistakes is

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

debatable. Bearing in mind that the pupils were also aged 6 and 7, tiredness and attention span would also be an issue.

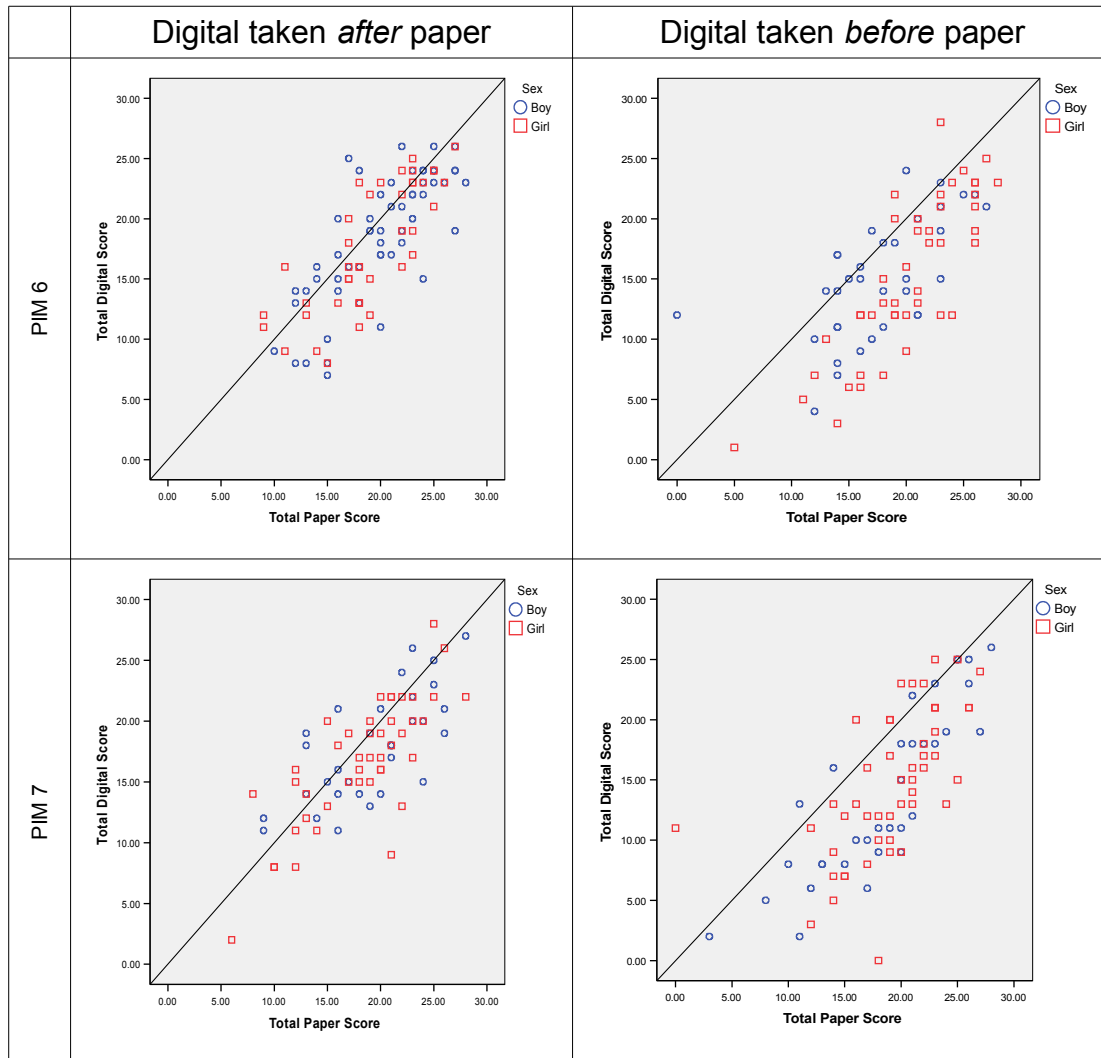


Figure 4.7: PIM Equating study - digital vs. paper grouped by order of testing

Was there a task effect?

Figures 4.9-4.16 show the average percentage score on each task at PIM6, PIM 7 and PIM 11 for pairs of digital and paper items. The lines joining the points have no real significance, but help make the graph readable. For the same reason, items in all of the charts are sorted according to the average facility of the paper version over all schools.

Note that these exclude the digital items that have no direct equivalent on paper – but include a few (such as “Fall” on PIM 6) where there is a non-identical paper item with obviously similar assessment objectives.

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

For each of the PIM 6 and PIM 7 tests, graphs are shown for the whole sample (Figs. 4.9,4.8), the students taking the paper test first (4.12,4.13) and students taking the digital test first (4.10,4.14). Figures 4.11 and 4.15 compare just the paper scores of the “paper first” and “paper second” groups, to see if having already seen the questions on computer affected the scores on paper.

For PIM 11, no order-of testing information was available, so a single graph showing overall digital vs. paper item scores is given (Fig. 4.16).

These graphs give a qualitative indication of how the difficulty of each “digital” item correlates with its paper equivalent, and which items contribute disproportionately to any overall difference. If the lines are consistently widely separated, it suggests that there is a systematic difference in difficulty affecting many questions; where the lines are jagged, questions are changing difficulty relative to other questions. However – it should be borne in mind that the number of cases varies between these plots - higher numbers will have a smoothing effect - the differences observed here will be tested for statistical significance later. As a rough guide differences of less than 10% (1 grid square) may not be significant. Note also that the numbers who took the digital test first are particularly low and (as observed above) represent students at just two schools.

The following things are apparent:

- By and large, the *relative* difficulty of the items is similar, suggesting that most of the questions are “testing the same thing”.
- A few items stand out as having larger discrepancies – several of these (such as “clock” on PIM7) correspond to issues noted later in the design analysis and observations.
- Discarding the schools who took the digital test first, or on the same day clearly eliminates a large source of discrepancy – leaving a few “problem” items.
- From Figs. 4.11 and 4.15, students taking the paper test seemed to perform similarly, regardless of whether they had already taken the digital test – and the items where there *was* a difference do not reliably correspond with items showing large paper/digital differences. If there was a school-based factor in the “digital first” schools it seems to be related to the administration of the digital test – possibly at school ID 5213 which dominates the “digital first” scores.
- The PIM 11 results appear more stable with only a few items showing notable discrepancies, and some cases of higher facility on digital versions.

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

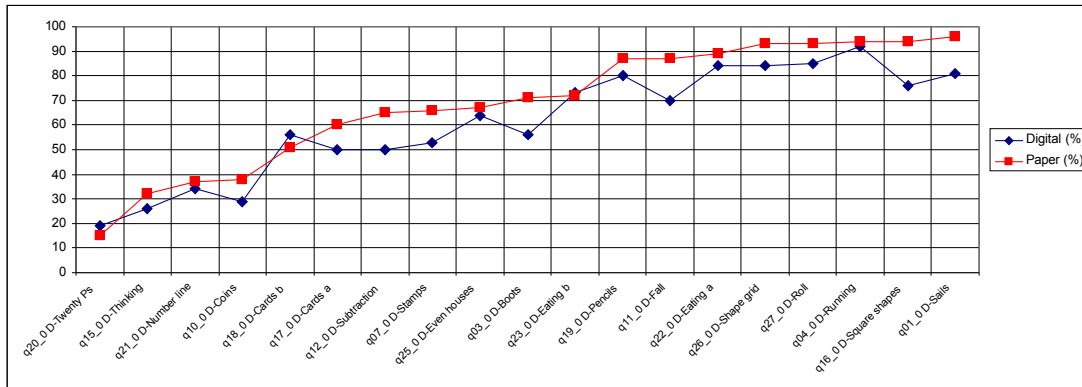


Figure 4.8: PIM 6 Equating study - digital vs. paper question facility levels - all schools N=181

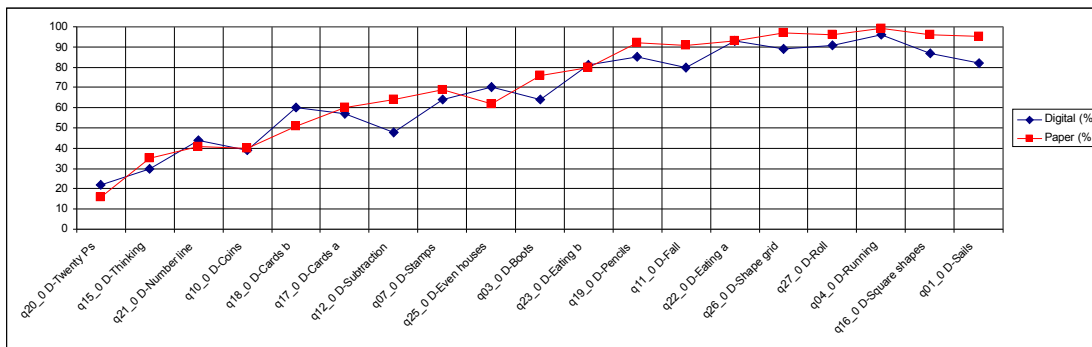


Figure 4.9: PIM 6 Equating study - digital vs. paper question facility levels - schools taking **paper** test first (N=97)

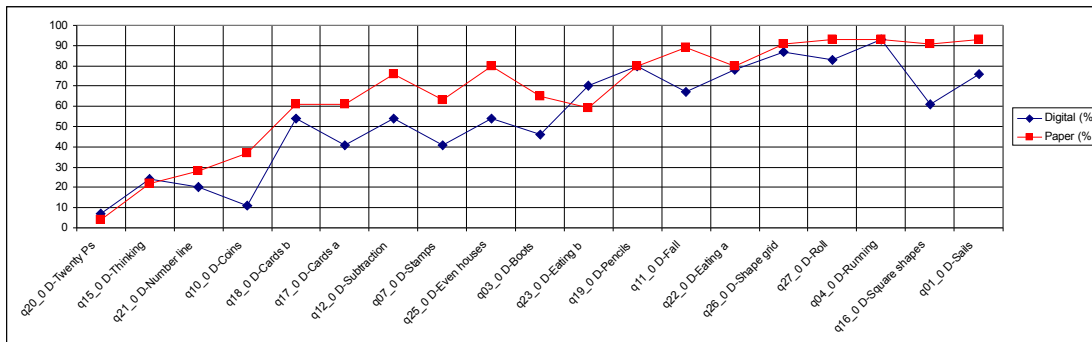


Figure 4.10: PIM 6 Equating study - digital vs. paper question facility levels - schools taking **digital** test first (N=46)

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

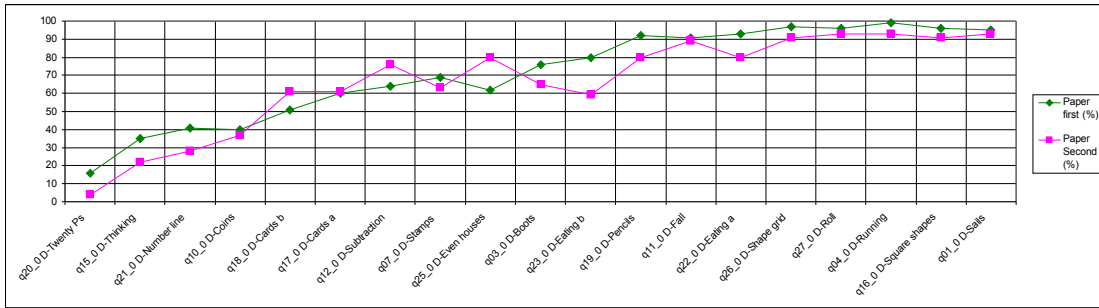


Figure 4.11: PIM 6 Equating study - paper question facility levels vs. order of testing (N=97 for paper first vs. N=46 for paper second)

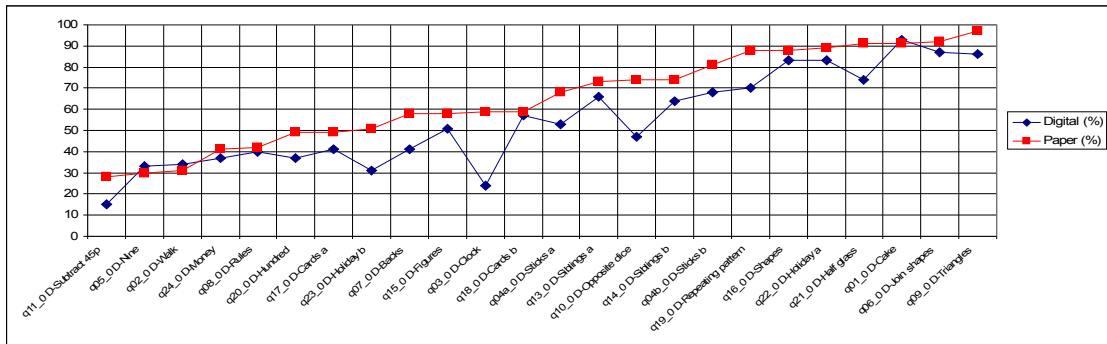


Figure 4.12: PIM 7 Equating study - digital vs. paper question facility levels - all schools (N=160)

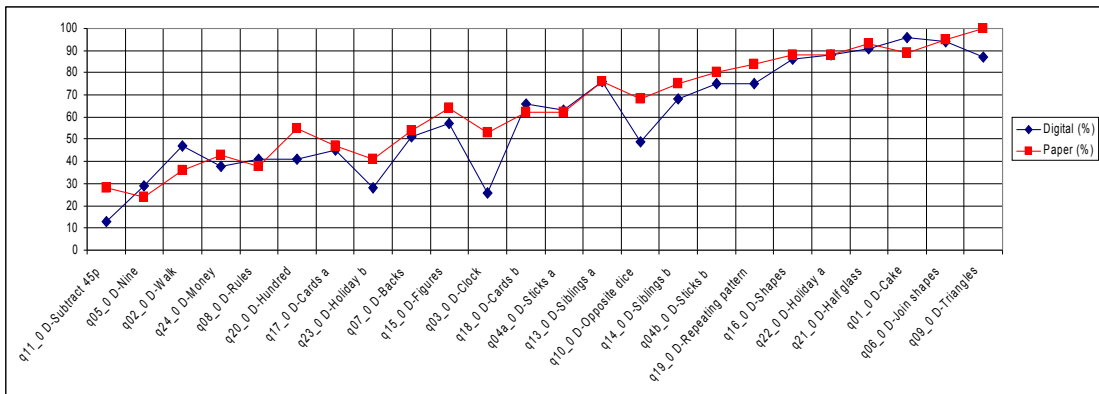


Figure 4.13: PIM 7 Equating study - digital vs. paper question facility levels - schools taking paper test first (N=76)

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

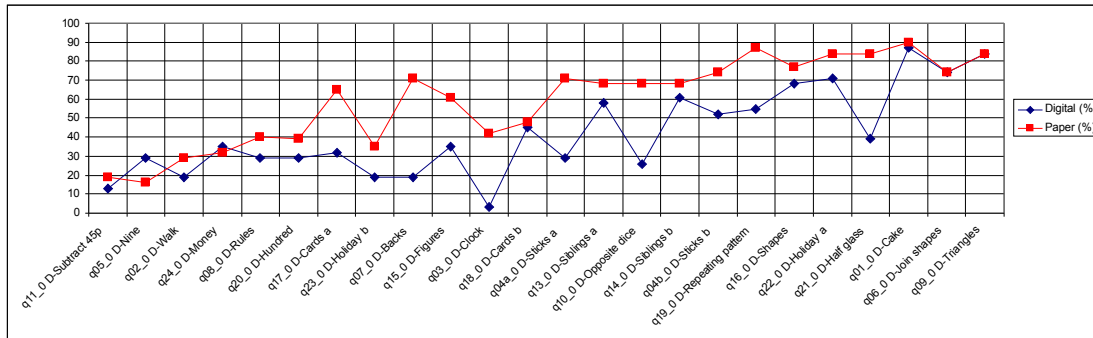


Figure 4.14: PIM 7 Equating study - digital vs. paper question facility levels - schools taking digital test first (N=31)

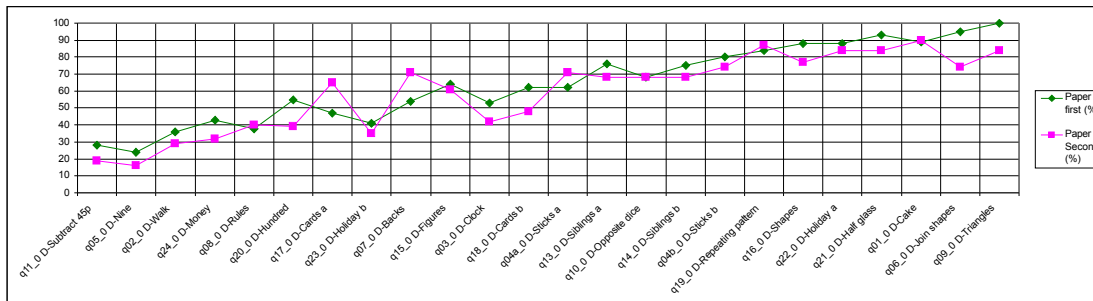


Figure 4.15: PIM 7 Equating study - paper question facility levels vs. order of testing (N= 76 for paper first vs. N=31 for paper second)

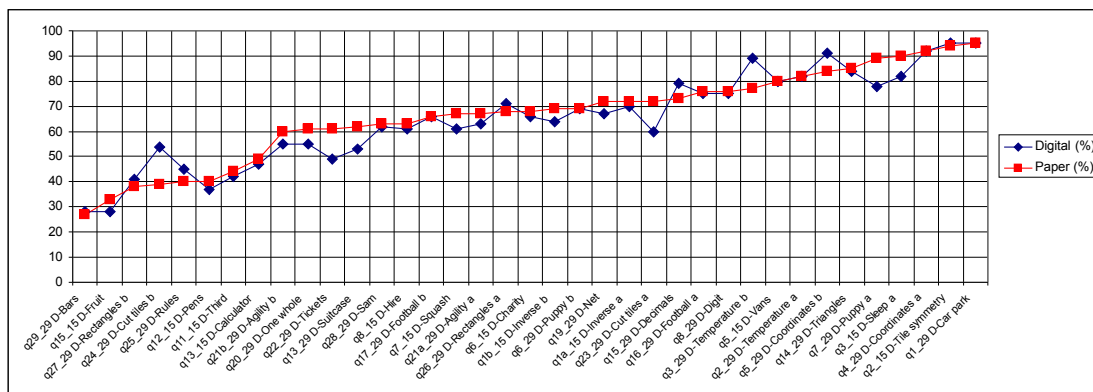


Figure 4.16: PIM 11 Equating study - digital vs. paper question facility levels - all schools (N=238)

Statistical tests of item differences

The following tables summarise the observed differences between digital and paper performance in the equating study. The data comprises:

- **Digital %**
The facility level (%) of the digital item
- **Diff %**
The difference in facility between the digital item and the equivalent paper item. A negative value means that the facility on the digital item was lower. Values shown in **bold** are those highlighted by nferNelson as being of concern.
- **Rasch**
The Rasch difficulty measure (Bond & Fox, 2001). This give an indication of the *relative* difficulty of each item (the larger the value, the more difficult) based on the equating study results, treating the paper and digital tests as a single test. - here it is being used as another significance test for the purported differences in facility (see also Pead, 2006, p. 41).
- **Rasch Shift**
The distance on the Rasch scale between the digital item and the paper equivalent. Negative values indicate that the digital item appeared “harder”. The advantage of Rasch here is that it reflects how the item discriminated between individual students in the sample relative to the other items, and it should be less sensitive to the composition of the sample than a simple mean score. Entries are in **bold** if this is more than twice the combined standard error for the Rasch measures of the items.
- **McNemar’s Test**
Typically, a chi-squared test could be used to determine whether the proportion of students passing or failing each item was dependent on which of the two versions of that item they took. However, this assumes that the two versions were given to separate samples – in this case, the same group of pupils took both versions, so chi-squared would be invalid. McNemar's test is similar to chi-squared, but only considers the students who have “passed” one version of the item and “failed” the other, and is valid for use on a single sample. Entries in the p column are bold for $p < 0.05$. **However**, this should be treated with caution: since we are individually testing 20-25 questions, 95% confidence means that a few results with $p \approx 0.05$ might be expected to occur by chance.

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

For items worth 2 marks, a “pass” is taken as full marks (such items offer 1 mark for partial credit – those questions with two independently awarded marks have been treated as two items).

It can be seen from the PIM 6 and 7 tables that the Rasch and McNemar methods largely support Nelson's identification of problem items, and that a limited number of questions do show significant differences in performance between digital and paper versions which warrant further investigation. At ages 6 and 7 it is almost always the digital version which shows the lowest score, but at age 11 there are differences in both directions.

For PIM 6 and 7, the McNemar results are also shown for the subset of pupils who took the digital test after the paper test (Tables 4.4 & 4.6) – this can be seen to reduce the number of questions with significant deviations. (This analysis was not done for pupils taking the digital test first since, as noted above, that part of the sample only represented 1 or 2 schools).

McNemar's Test	
Pf/Df	Number failing both paper & digital
Pf/Dp	Number failing paper & passing digital
Pp/Df	Number passing paper & failing digital
Pp/Dp	Number passing both paper and digital
Colour coding:	
Light grey	No significant difference
White	Significant digital underperformance (by McNemar's test)
Dark grey	Significant digital over-performance (by McNemar's test)

Figures in **bold** satisfy the relevant significance test.

Table 4.2: Key to statistics tables

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

Qn	Item	Facility		Rasch		McNemar's Test				Total	χ^2	p
		Digital %	Diff %	Difficulty	shift	Pf/Df	Pf/Dp	Pp/Df	Pp/Dp			
q01_0	Sails	81	-15	-0.80	-1.88	6	2	29	144	181	21.807	0.000
q03_0	Boots	56	-15	0.69	-0.86	42	10	38	91	181	15.188	0.000
q04_0	Running	92	-2	-2.02	-0.29	5	6	9	161	181	0.267	0.606
q07_0	Stamps	53	-13	0.84	-0.72	48	13	37	83	181	10.580	0.001
q10_0	Coins	29	-9	2.18	-0.52	100	13	29	39	181	5.357	0.021
q11_0	Fall*	70	-17	-0.07	-1.27	14	10	41	116	181	17.647	0.000
q12_0	Subtraction	50	-15	1.01	-0.82	49	14	42	76	181	13.018	0.000
q15_0	Thinking	26	-6	2.35	-0.37	100	23	34	24	181	1.754	0.185
q16_0	Square shapes	76	-18	-0.49	-1.82	7	4	36	134	181	24.025	0.000
q17_0	Cards a	50	-10	1.01	-0.52	53	20	38	70	181	4.983	0.026
q18_0	Cards b	56	4	0.69	0.24	53	35	27	66	181	0.790	0.374
q19_0	Pencils	80	-8	-0.72	-0.67	15	8	22	136	181	5.633	0.018
q20_0	Twenty Ps	19	4	2.87	0.33	137	17	10	17	181	1.333	0.248
q21_0	Number line	34	-3	1.85	-0.16	95	19	24	43	181	0.372	0.542
q22_0	Eating a	84	-5	-1.08	-0.49	5	15	24	137	181	1.641	0.200
q23_0	Eating b	73	1	-0.31	0.07	20	30	28	103	181	0.017	0.896
q25_0	Even houses	64	-3	0.25	-0.19	38	21	27	95	181	0.521	0.470
q26_0	Shape grid	84	-9	-1.08	-1.13	6	6	23	146	181	8.828	0.003
q27_0	Roll	85	-8	-1.18	-1.88	7	5	20	149	181	7.840	0.005
	Mean		-7.7		-0.68							
	Median		-8		-0.52							

Table 4.3: PIM6 - Whole equating sample

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

Qn	Item	Facility		McNemar's Test						
		Digital %	Diff %	Pf/Df	Pf/Dp	Pp/Df	Pp/Dp	Total	χ^2	p
q01_0	Sails	82	-12	4	1	13	79	97	8.6429	0.00328
q03_0	Boots	64	-12	16	7	19	55	97	4.6538	0.03098
q04_0	Running	96	-3	1	0	3	93	97	1.3333	0.24821
q07_0	Stamps	64	-5	22	8	13	54	97	0.7619	0.38273
q10_0	Coins	39	-1	47	11	12	27	97	0.0000	1.00000
q11_0	Fall*	80	-10	3	6	16	72	97	3.6818	0.05501
q12_0	Subtraction	48	-15	25	10	25	37	97	5.6000	0.01796
q15_0	Thinking	30	-5	47	16	21	13	97	0.4324	0.51080
q16_0	Square shapes	87	-9	2	2	11	82	97	4.9231	0.02650
q17_0	Cards a	57	-3	28	11	14	44	97	0.1600	0.68916
q18_0	Cards b	60	9	29	19	10	39	97	2.2069	0.13739
q19_0	Pencils	85	-7	6	2	9	80	97	3.2727	0.07044
q20_0	Twenty Ps	22	5	69	12	7	9	97	0.8421	0.35880
q21_0	Number line	44	3	42	15	12	28	97	0.1481	0.70031
q22_0	Eating a	93	0	1	6	6	84	97	0.0833	0.77283
q23_0	Eating b	81	1	7	12	11	67	97	0.0000	1.00000
q25_0	Even houses	70	8	21	16	8	52	97	2.0417	0.15304
q26_0	Shape grid	89	-8	0	3	11	83	97	3.5000	0.06137
q27_0	Roll	91	-5	2	2	7	86	97	1.7778	0.18242
	Mean		-3.6							
	Median		-5							

Table 4.4: **PIM6** - Students taking digital test after paper test

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

Qn	Item	Facility		Rasch		McNemar's Test						
		Digital %	Diff %	Difficulty	Shift	Pf/Df	Pf/Dp	Pp/Df	Pp/Dp	Total	χ^2	p
q01_0	Cake	93	2	-2.38	0.27	2	13	10	135	160	0.174	0.677
q02_0	Walk	34	3	1.48	0.18	87	24	19	30	160	0.372	0.542
q03_0	Clock	24	-35	2.06	-1.91	60	5	61	34	160	45.833	0.000
q04a_0	Sticks a	53	-14	0.47	-0.75	31	21	44	64	160	7.446	0.006
q04b_0	Sticks b	68	-13	-0.28	-0.85	18	13	34	95	160	8.511	0.004
q05_0	Nine	33	3	1.52	0.18	89	23	18	30	160	0.390	0.532
q06_0	Join shapes	87	-5	-1.13	-0.26	10	5	17	128	160	5.500	0.019
q07_0	Backs	41	-16	1.08	-0.83	53	15	41	51	160	11.161	0.001
q08_0	Rules*	40	-2	1.15	-0.15	81	35	15	29	160	7.220	0.007
q09_0	Triangles	86	-11	-1.61	-1.78	2	3	20	135	160	11.130	0.001
q10_0	Opposite dice	47	-27	0.79	-1.45	33	9	52	66	160	28.918	0.000
q11_0	Subtract 45p	15	-13	2.79	-0.97	112	3	24	21	160	14.815	0.000
q13_0	Siblings a	66	-7	-0.21	-0.41	33	10	21	96	160	3.226	0.072
q14_0	Siblings b	64	-10	-0.11	-0.58	29	12	28	91	160	5.625	0.018
q15_0	Figures	51	-8	0.6	-0.38	53	14	26	67	160	3.025	0.082
q16_0	Shapes	83	-5	-0.81	-0.35	16	8	18	118	160	3.115	0.078
q17_0	Cards a	41	-9	1.11	-0.45	66	15	29	50	160	3.841	0.050
q18_0	Cards b	57	-2	0.28	-0.09	46	20	23	71	160	0.093	0.760
q19_0	Repeating pattern*	70	-18	-0.43	-1.31	8	12	40	100	160	14.019	0.000
q20_0	Hundred	37	-13	1.31	-0.65	68	13	33	46	160	7.848	0.005
q21_0	Half glass	74	-16	-0.69	-1.42	9	6	32	113	160	16.447	0.000
q22_0	Holiday a	83	-7	-1.28	-0.67	7	10	21	122	160	3.226	0.072
q23_0	Holiday b	31	-21	1.66	-1.09	66	12	45	37	160	17.965	0.000
q24_0	Money	37	-4	1.31	-0.2	74	21	27	38	160	0.521	0.470
	Mean		-10		-0.66							
	Median		-9.5		-0.62							

Table 4.5: PIM7 - Whole equating sample

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

Qn	Item	Facility Digital %	Diff %	McNemar's test						
				Pf/Df	Pf/Dp	Pp/Df	Pp/Dp	Total	χ^2	p
q01_0	Cake	96	7	0	8	3	65	76	1.455	0.228
q02_0	Walk	47	12	32	17	8	19	76	2.560	0.110
q03_0	Clock	26	-26	33	3	23	17	76	13.885	0.000
q04a_0	Sticks a	63	1	14	15	14	33	76	0.000	1.000
q04b_0	Sticks b	75	-5	8	7	11	50	76	0.500	0.480
q05_0	Nine	29	5	47	11	7	11	76	0.500	0.480
q06_0	Join shapes	94	-1	1	4	5	66	76	0.000	1.000
q07_0	Backs	51	-3	27	8	10	31	76	0.056	0.814
q08_0	Rules*	41	3	42	16	3	15	76	7.579	0.006
q09_0	Triangles	87	-13	0	0	10	66	76	8.100	0.004
q10_0	Opposite dice	49	-20	16	8	23	29	76	6.323	0.012
q11_0	Subtract 45p	13	-14	53	2	13	8	76	6.667	0.010
q13_0	Siblings a	76	0	12	6	6	52	76	0.083	0.773
q14_0	Siblings b	68	-7	14	5	10	47	76	1.067	0.302
q15_0	Figures	57	-8	19	8	14	35	76	1.136	0.286
q16_0	Shapes	86	-2	4	6	9	57	76	0.267	0.606
q17_0	Cards a	45	-3	34	6	8	28	76	0.071	0.789
q18_0	Cards b	66	4	18	11	8	39	76	0.211	0.646
q19_0	Repeating pattern*	75	-9	4	8	15	49	76	1.565	0.211
q20_0	Hundred	41	-14	27	7	18	24	76	4.000	0.046
q21_0	Half glass	91	-3	1	4	6	65	76	0.100	0.752
q22_0	Holiday a	88	0	4	5	5	62	76	0.100	0.752
q23_0	Holiday b	28	-13	39	6	16	15	76	3.682	0.055
q24_0	Money	38	-5	33	10	14	19	76	0.375	0.540
	Mean		-4.8							
	Median		-3							

Table 4.6: PIM7 - Students taking digital after paper

* Questions marked with an asterisk are new, digital-only tasks which appear to be a “replacement” for a particular paper task, but which are clearly not the same question. While it is interesting to note whether the replacement questions are comparable in difficulty to the originals, there is no reason to expect them to be equivalent.

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

Qn	Item	Digital %	Diff %	Rasch		McNemar's Test						
				Diffi- culty	Shift	Pf/D f	Pf/Dp	Pp/Df	Pp/Dp	Total	χ^2	p
q1a_15	Inverse a	70	-2	-0.14	-0.1	45	22	26	145	238	0.188	0.665
q1b_15	Inverse b	64	-5	0.19	-0.28	49	24	36	129	238	2.017	0.156
q2_15	Tile symmetry	95	1	-2.52	0.17	5	9	7	217	238	0.063	0.803
q3_15	Sleep a	82	-8	-0.93	0.93	13	10	30	185	238	9.025	0.003
q5_15	Vans	80	0	-0.8		25	22	22	169	238	0.023	0.880
q6_15	Charity	66	-2	0.08	-0.12	58	17	22	141	238	0.410	0.522
q7_15	Squash	61	-6	0.35	-0.32	60	18	32	128	238	3.380	0.066
q8_15	Hire	61	-2	0.37	-0.09	72	17	21	128	238	0.237	0.627
q11_15	Third	42	-3	1.25	-0.11	129	19	27	63	238	1.065	0.302
q12_15	Pens	37	-4	1.63	-0.21	123	19	28	68	238	1.362	0.243
q13_15	Calculator	47	-2	1.08	-0.09	92	30	34	82	238	0.141	0.708
q15_15	Fruit	28	-5	2.14	-0.3	142	18	30	48	238	2.521	0.112
q1_29	Car park	95	-1	-2.33	0.18	4	15	20	199	238	0.457	0.499
q2_29	Temperature a	82	-1	-0.89	-0.07	24	18	20	176	238	0.026	0.871
q3_29	Temperature b	89	12	-1.6	1.04	14	41	12	171	238	14.79	0.000
q4_29	Coordinates a	92	0	-1.99	0	3	16	16	203	238	0.031	0.860
q5_29	Coordinates b	91	6	-1.81	0.67	11	26	11	190	238	5.297	0.021
q6_29	Puppy b	69	0	-0.09	0.02	39	35	34	130	238	0.000	1.000
q7_29	Puppy a	78	-11	-0.62	-0.98	7	19	46	166	238	10.4	0.001
q8_29	Digit	75	-1	-0.45	-0.08	44	12	15	167	238	0.148	0.700
q13_29	Suitcase	53	-9	0.76	-0.46	68	22	43	105	238	6.154	0.013
q14_29	Triangles	84	-1	-1.1	-0.08	23	13	15	187	238	0.036	0.850
q15_29	Decimals	79	6	-0.71	0.42	40	25	10	163	238	5.600	0.018
q16_29	Football a	75	-1	-0.45	-0.08	28	28	31	151	238	0.068	0.795
q17_29	Football b	66	0	0.1	0	56	25	25	132	238	0.020	0.888
q19_29	Net	67	-5	0.05	-0.29	44	23	35	136	238	2.086	0.149
q20_29	One whole	55	-6	0.71	-0.25	94	16	38	90	238	8.167	0.004
q21a_29	Agility a	63	-4	0.26	-0.23	54	24	34	126	238	1.397	0.237
q21b_29	Agility b	55	-5	0.69	-0.25	70	26	38	104	238	1.891	0.169
q22_29	Tickets	49	-11	0.97	-0.58	71	23	50	94	238	9.260	0.002
q23_29	Cut tiles a	60	-13	0.44	-0.7	55	11	41	131	238	16.173	0.000
q24_29	Cut tiles b	54	15	0.72	0.77	88	57	21	72	238	15.705	0.000
q25_29	Rules	45	5	1.16	0.26	96	46	34	62	238	1.513	0.219
q26_29	Rectangles a	71	3	-0.21	0.19	43	33	25	137	238	0.845	0.358
q27_29	Rectangles b	41	3	1.4	0.14	121	26	20	71	238	0.544	0.461
q28_29	Sam	62	-2	0.33	-0.09	60	27	31	120	238	0.155	0.694
q29_29	Bars	28	1	2.14	0.05	153	21	19	45	238	0.025	0.874
	Mean		-1.6		-0.03							
	Median		-2		-0.085							

Table 4.7: PIM 11 - Whole equating sample

Use of advanced scaling techniques

One test applied to these items was to use *Rasch scaling* (see Bond & Fox, 2001). This technique, based on Item Response Theory, places test items on a relative difficulty scale using a probabilistic model – if two items are of equivalent difficulty, the odds of a particular candidate passing them should be the same. Because this is based on the relative discrimination of items, the scale positions of the items should be largely independent of the composition of the sample.

Rasch analysis can be used for test calibration and anchoring, or “reversed” and used to assign an ability scale to the candidates. Here, the scores of the equating study group on both

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

the digital and paper versions of a test were combined and analysed as a single group, to test the assertion that the digital items were equivalent in difficulty to the paper versions. The analysis process produces a standard error for the difficulty – pairings of digital and paper questions for which the measure differed by more than two standard errors are highlighted in the tables above.

The Rasch model makes some quite important assumptions, in particular that of unidimensionality: it is presumed that all the questions are measuring the same quantity, and that the candidates possess varying levels of that same quantity. It would not fit if, for example, some questions were measuring a completely different skill to others, or if some candidates had been taught a fundamentally different curriculum to others. Although this unidimensionality may appear simplistic, it should be remembered that the same assumption is implicit in any assessment which aims to produce a single, summative score for each candidate.

One output of the Rasch analysis is a measure of “fit” for each item and candidate: in theory, this should highlight items which fail to fit the Rasch model and which may, therefore, be mis-performing. If the “fit” for an item is too low (“underfit”), this indicates that it is producing random, noisy results uncorrelated with the rest of the items (so, for example, pupils may be guessing the answer, or it might be measuring something unrelated to the other questions). Too high a value (“overfit”) is harder to reconcile with features of the question design, but relates to the probabilistic nature of the Rasch model, and suggests that pupils’ performances on the item are too predictable, with no “intermediate” cases.

In addition to the scaling of the combined digital and paper results, the individual tests at ages 6 and 7 were scaled separately, to see if the Rasch fit measures would identify potential problem items (Figs. 4.17 & 4.18¹¹). The usual criteria for determining whether an item fits the Rasch model is an infit statistic between 0.77 and 1.3 – the figures show that all of the Age 6 items and most of the Age 7 items meet these criteria. The single case of “underfit” (“CardsB” at Age 7) does correspond to a question identified in the design critique as amenable to guessing – but it would be dangerous to draw any conclusions from this single case – especially as the “guessing” criticism applies equally to the paper version which did not show underfit.

Conversely, tasks which were clearly identified as “not testing mathematics” in the design critique and school observation – “Clock” at ages 6 and 7 and “Rules” at age 7 – show an acceptable fit statistic. Consequently, there is no evidence here that the Rasch “fit” statistic can be reliably used to detect mis-performing tasks.

11 This analysis includes a few completely new digital tasks which were omitted from the earlier comparison as they had no paper equivalent.

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

- If the raw data are correct, some schools took the digital test on the same day as the digital test. It is not possible to say in which order these pupils took the tests, and hard to predict what the effect might be on 6 or 7 year-old students of taking two very similar (but not identical) sets of questions so close together.
- Since the two tests contain near-identical questions some sort of repeat testing effect could be anticipated – but prior knowledge could equally well help or hinder if answers were misremembered or, as in a few cases, the two versions had different answers. The sample is not well balanced between schools who took the computer test first and those who took it after the paper: while 30-40 students took the digital test first, they are predominantly from a single school, so the possibility of a school-wide effect makes it unsafe to quantify any test/retest effect from this data.
- There is a strong correspondence between the few schools taking the digital test first and the schools showing large discrepancies. However, comparing their paper scores with other schools does not suggest a significant difference in ability or any huge advantage in having seen the test before. It is possible that there was a problem with the administration or organisation of the digital tests in these schools.
- The schools taking the digital test after the paper test show a more consistent performance, with just a few items showing significant discrepancies (especially considering that one or two results out of 20 with $p \approx 0.05$ would not be significant). Whether there is justification for selectively analysing this subset of the results is debatable – but tables 4.4 and 4.6 show what the effect would be.
- The study does support the notion that certain questions showed a marked difference in performance compared with others. The Rasch test, in particular, indicates changes in the **relative** difficulty of tasks as experienced by the sample, and should be fairly insensitive to order-of-testing and school-wide issues. It does, therefore, appear that a subset of the questions at ages 6 and 7 are more difficult in their digital form.
- For PIM11 data on the order of testing was not available, but the results for the whole sample show that significant discrepancies are confined to a few items. Some items appear easier on the digital test (which was rare with PIM 6 and PIM 7).
- One pattern emerging at PIM11 is that the second parts of several 2-part questions show significantly higher facilities on the digital test. This could be due to changes in the question, but it is possible that re-stating the question on a fresh page has an effect: on paper, students may habitually skip the second part of a question if they have difficulty with the first part, whereas the computer presentation might encourage them to treat it as a fresh question.

The design of the original equating study limits the conclusions that can be drawn from this data. Having all candidates take both versions of the test convolutes any digital vs. paper effect with a test/re-test effect. Since no attempt was made to control order of testing or the interval between tests, the test/retest effect would be difficult to eliminate.

One alternative would have been to use a “cross-over” study in which half the sample took the paper test, and the other half the digital version – although the experience of attempting this in the GCSE study (see Chapter 6) is that a large sample of schools would be needed to ensure a balanced sample. In this case, the performance of the paper test alone was well known from prior, large-scale calibration exercises, so a better approach might have been to have a sample of students take just the digital test, and compare their performance with the paper data.

4.5: Task design critique & school observations

The full commentary on the tasks and school observations is can be found in the report to nferNelson (Pead, 2006) includes details of the data analysis, design critique and school observation on a task-by-task basis.

Here we summarise features of the design of both the overall testing system and individual tasks which were identified as possible sources of difficulty during the critique or observations.

Presentation of spoken prompts

In the established paper tests for ages 6-8, each question “prompt” was read out by the teacher according to a script in the teacher's guide. The question in the pupils workbook did not include this text, and many questions could not be answered without first listening to the prompt.

The design of the computer tasks was essentially the same, with the question posed to the pupil by a recorded voice.

One obvious issue raised by this was whether the use of a recorded voice could be assumed to be equivalent to having the pupils' own teacher read the questions. To explore this, the teachers of the observed pupils were asked whether they have previously administered PIM 6 or 7 paper tests, and if so, how they went about presenting a task (see Pead, 2006, pp. 34-35). It was evident from this that typical teachers' actions included some or all of the following:

- Ensuring students are paying attention and looking at the correct page

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

- Pointing to parts of the question
- Reading the prompt at least twice
- Getting the students to start work on the task
- Stopping the students after a spell, and reading the prompt again

Compare this with the digital tests in which:

- The screen is displayed
- Immediately, the voice begins to read out the whole prompt
- The voice continues regardless of what the students are doing (e.g. if they have started working out the answer, or have answered the first part of a two-part question)
- The prompts are often longer than paper equivalents, sometimes because of added computer operation instructions (click here, move the slider etc.)
- The prompt is only repeated if the student chooses to click on the “Listen again” button

It seems unreasonable to assume that these two modes are equivalent. Even disregarding any differences in the clarity or familiarity of the voice, it is clear that most teachers were proactive in ensuring that their pupils were paying attention to the question. There is also the indirect effect that, whereas the digital tests were self-paced, pupils would work through the paper tests in step, with the pacing controlled by their teacher.

During the observations, students were often seen not paying attention to the prompt, rushing to a wrong answer and moving on. They often corrected themselves when the observer intervened by instructing them to try “listening again”. It was rare for pupils to “listen again” spontaneously: contrast this with teachers' practice of habitually reading out each question two or three times.

When not paying attention to the prompts, children were liable to spot a “question” on the screen (e.g. count the studs on the football boot) and answer that regardless of the prompt (how many studs on a **pair** of boots). For example, the following interaction was observed:

Computer: “Maya wants to post **ten cards**. She needs **one stamp** for **each card**. But she has only **six stamps**. **How many more stamps** does she need?”

Pupil: Counts the post-cards or stamps in the picture – enters 10 or 6 in the answer box and moves on.

The timing of the recorded voice was sometimes a problem: instances were observed of pupils starting to engage with the question during a pause in the speech, only to be distracted

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

when the voice continued. Where questions included instructions such as “Click and type your answer...” after the main question, or, in questions using illustrative animations, “Click the green button to see (the object) move again...” some students were observed to act on these immediately, interrupting their thinking or, in the case of the animation, causing it to start again before they had watched it all the way through.

Other prompts – such as the opening narration – appeared to be too long to hold the attention of younger pupils, who were expected to remember instructions for later in the test, with no visual or “try it now” reinforcement.

Another notable difference between the two media was the level of discretion given to teachers to re-phrase the question prompts for the paper test, compared with the fixed, pre-recorded prompts used by the digital test. Instances were observed in which pupils taking the digital test had problems with vocabulary, and the substitution of (for example) “numbers” for “figures” or “more than” for “greater than” by the observer allowed them to continue with the question.

The teacher's instructions for the paper PIM test provide a script for each question, but state “The wording of these questions may be adapted, provided the meaning is retained.” and also inform teachers that “you may explain any non-mathematical words or expressions, but you should not help with the mathematical content of individual questions” (Clausen-May et al., 2004a). No detailed question-by-question guidance on how to draw this distinction is given apart from one example per test of a term which shouldn't be “explained” (there are no counter examples showing acceptable substitutions). It is therefore up to the teacher to decide whether (for example) substituting the word “numbers” for “figures” constituted “explaining a mathematical term” in the context of 6-7 year old pupils.

The teachers interviewed were unanimous that they would not explain a **mathematical** term to students – but some were prepared to “use their own words” or “explain words and phrases” so whether the above substitutions were made would be at the discretion of the teacher administering the paper test¹².

Arguably, the use of a recorded voice should produce more reliable test results by removing teachers' discretion and presenting the prompts in a consistent manner. Conversely, deliberate action by the teacher to maintain concentration and ensure that pupils were answering the intended question could help ensure that the test was working as intended and measuring mathematical ability rather than language skills or attention span.

While the spoken questions make the tests accessible to children with with low reading ability, it was apparent from the observations and interviews that less-able children in these

¹² This is mentioned here purely as a possible source for a digital vs. paper discrepancy: no criticism of these teachers' actions is intended and it should be noted that the PIM tests are not “high stakes” tests.

age groups have deeper issues with language comprehension, communication and concentration skills. The tendency for computer prompts to be longer than the paper equivalents (to accommodate operating instructions) and mechanically delivered (where a teacher would be aware of the pupils' attention) is a potential problem with computer-based tests at this age.

Pictures and distraction

Research has shown that care needs to be taken when illustrating mathematics questions (Santos-Bernard, 1997). If pupils are intended to extract information from the illustration (for example by counting objects) some might ignore the illustration and use numbers from the question text. Conversely, if all the information required is contained in the question text, but a purely illustrative illustration contains contradictory data or suggests a different question, some students will ignore the text in favour of the illustration.

The paper versions of the PIM tests already made extensive use of illustrations. The digital versions are, mostly, closely based on these, with the addition of full colour. In some cases, minor changes to layout had been made, and the detailed commentary on the tasks notes some cases where this might have had an effect on the response.

However, the observations also found evidence of pupils being distracted by illustrations displayed while the test instructions were being played.

One such screen is shown in Fig. 4.19 – one of the first screens encountered by children taking the test. At this point, the recorded voice is telling the pupil that they are about to take a maths test; that the questions will be read out to them; that they can hear them again by clicking on the “Listen again” button and that they should use the “Next” button to move on to the next screen.

Out of 34 pupil pairs observed encountering this screen, 8 requested or required intervention from the observer before they pressed “next” and continued, while a further 5 appeared to show some signs of distraction.

A typical interaction was:

Child: *(looking at the screen shown in Fig. 4.19):* “What do you do”
Observer: “What do you think you have to do?”
Child: “I think you have to put the numbers in order.”

A later screen, also simply a background image displayed while spoken instructions were being played, showed a photograph of a tree. Again, pupils simply had to press “next” and again 8 out of 34 pairs required assistance, while a further 18 pairs appeared distracted (moving the mouse around the screen or making comments about the “tree”).

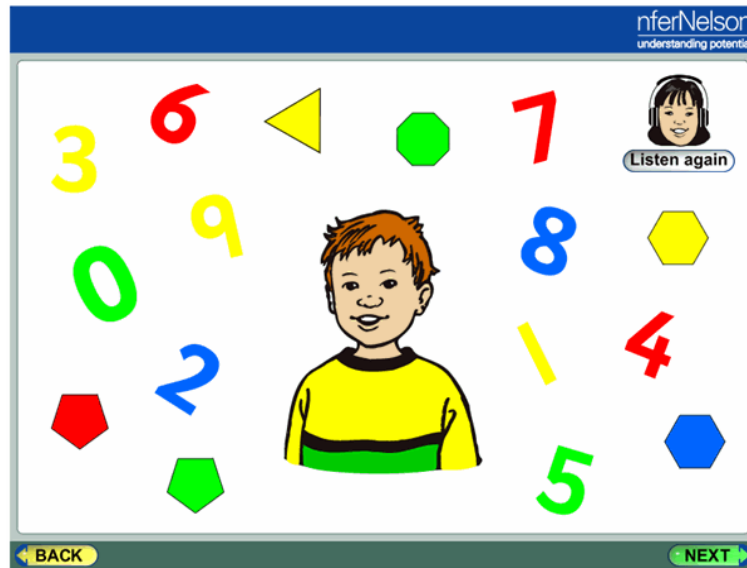


Figure 4.19: Background screen with irrelevant “mathematical” content

Neither of these screens had any direct bearing on the pupils' scores, and there is no evidence from the observations that their failure to pay attention to the instructions seriously impacted their performance. However, these observations do clearly show the potential of ill-chosen illustrations to cause distractions.

Colour and accessibility

Whereas the paper test booklets were each printed using a “two colour” process, the digital tests used a full range of colours. Consequently, several of the digital items introduced possible issues for pupils suffering from colour blindness – or poorly adjusted computer screens.

For example, Figure 4.20 requires a student to distinguish between blue and green shapes, otherwise they will need to apply some additional deductive reasoning in order to answer. Another question involved the continuation of a shape pattern: while the paper version was based entirely on shape the computer version had added colours. During the observations students were clearly heard referring to the shapes by colour rather than name – so it is feasible that colour-blind students would approach the task differently.

Even with perfect colour vision, colour display on monitors can be variable. In the case of flat-panel monitors, colours can vary considerably with viewing angle, and 6-year-old children (especially in a computer lab that also has to cater for older children) may not be looking from the optimum angle. One task in particular (“Rules” from PIM 7 – see Figure

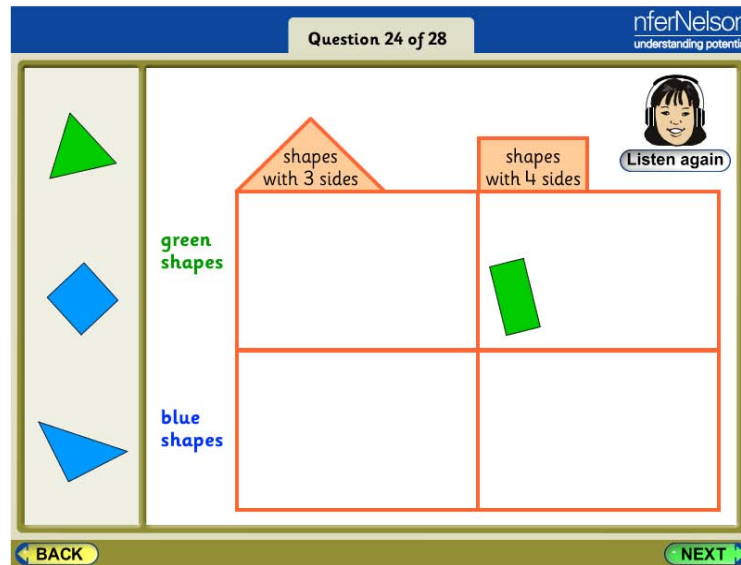


Figure 4.20: Colour and text visibility issues

4.31 later) was observed to cause problems with items such as “the blue numbers” and “the red box” referenced by the prompt being ambiguous on screen.

A proper investigation of the implications of these and other accessibility-related issues, while important, was beyond the scope of the evaluation. It is, however an example of how apparently simple design decisions which arise during the process of computerising a paper test, such as the use and choice of colours, can raise important new issues.


Changes in predicted difficulty

Generally, the digital versions were fairly faithful interpretations of the original paper tasks. A small number, however, had been altered to fit the structure of the digital test in a way which appeared to make them easier.

For example, while one paper question comprised two “number pyramid” questions, the digital version (Figure 4.21) featured the first part but not the second, more challenging part (Figure 4.22) which required subtraction, rather than addition as well as an extra step of reasoning. Although the overall facility level of the test may have been maintained by the presence of new questions of comparable difficulty, this represents a drift towards fragmentation and reduced “reasoning length” (see Section 3.2).

nferNelson
understanding potential

Question 5 of 31


[Listen again](#)

	10	
3	7	12

[← BACK](#)
[NEXT →](#)

Figure 4.21: Digital version of the “number pyramids” task (PIM 8). Each box must be the sum of the two boxes below - fill in the empty boxes

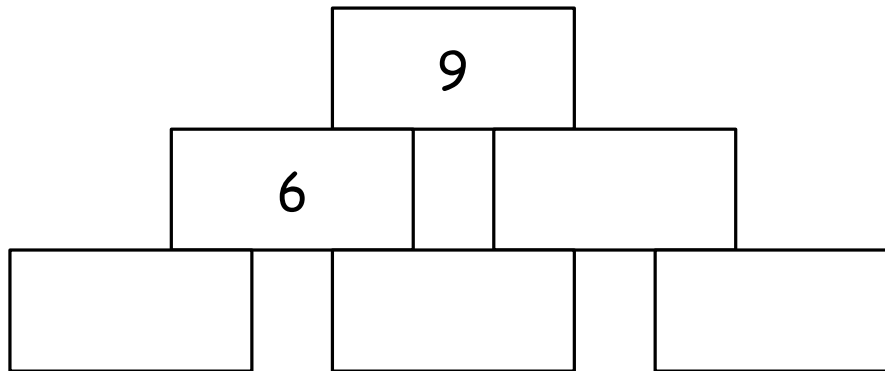


Figure 4.22: The second part of the paper version of the “number pyramid” task

nferNelson
understanding potential

Question 19 of 31



3

6

7

-


-

$3 + 3 = 6$ $8 - 2 = 6$

4 + 2 = 6

5 + 1 = 6



Listen again

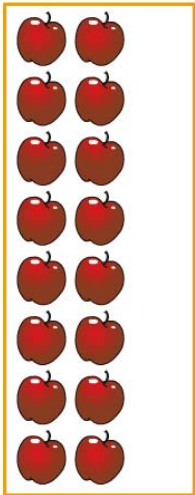
← BACK NEXT →

Figure 4.23: Constrained responses


nferNelson
understanding potential


Question 25 of 31

Maria



Suki





Listen again

- +

← BACK NEXT →

Figure 4.24: This can be solved by clicking “+” until the two piles look similar (PIM 8)

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

In a similar case, a paper question gave two examples of simple sums with the answer six, and asked for “three more ways of making six” - students could write down any three calculations they could think of. The digital version (Figure 4.23) reduced this to “find **two** more ways of making six” and introduced a drag-and-drop system that constrained the possible responses to addition/subtraction of two single-digit numbers, without re-using any digits.

In other cases, the computer enables alternative, possibly easier, methods – e.g. Figure 4.24 could easily be answered visually by adding or removing apples until the two piles are visibly the same - no counting or calculation required.

Splitting a question across two screens might increase the facility on the second part. The distraction of having the first question on the screen is removed, and students who were stuck on the first part may be more inclined to attempt the second part when it is presented separately. There is some suggestion of this in the equating study at PIM 11 (but often complicated by other changes in the question). This also illustrates that these longer tasks are simply separate questions with a common theme and do not represent a true “extended” task in which the second part builds on the work done in the first part.

Many of the instances of tasks possibly being made easier as a result of their translation to computer come from the 8 year-old-test – at which stage slightly more sophisticated questions were being introduced on paper but the on-screen user interface design was substantially the same as for the age 6 and 7 tests. By age 11, the screen design had evolved to include more information and text on each screen.

There was no equating study or observation data for age 8, so the effects predicted above could not be investigated further.

Validity of rich contexts and problem solving

Although the PIM tests (both digital and paper) clearly make an effort to present mathematics in realistic contexts wherever possible, none of the tasks fully meet the criteria for rich, open-ended, problem solving tasks featuring extended chains of reasoning discussed in Section 2.4.

However, richer tasks almost inevitably require more verbose questions and a significant extra comprehension step. This is a particular obstacle in the case of “formal” assessment of 5-7 year olds, whose comprehension skills (even when reading is eliminated) and life experience cover such a wide range. Of the pupils observed taking PIM, only the most able appeared to have the comprehension skills and attention span required to tackle extended

problems unaided in a test environment. It is to be hoped that pupils of this age will encounter such tasks in a more supportive and collaborative classroom environment ¹³.

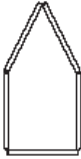
A related concern is that the conflict between the desire for “rich” questions and the constraints of comprehension, time and attention span can result in tasks that appear superficially rich but are effectively just a simple counting or arithmetic exercise presented as a word problem. Figure 4.25 shows part of an example of one *World Class Tests* task based on a well-known “problem solving” genre.

1.1

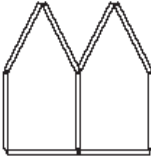
straw houses

generalising

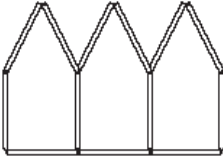
Lindsay uses drinking straws to make houses in a row.



1 house
5 straws



2 houses
9 straws



3 houses
13 straws

Five straws make one house, nine straws make two houses and thirteen straws make three houses.

1. Draw a diagram to show four houses in a row.

Lindsay makes a table to show the number of straws needed for rows with different numbers of houses.

Number of houses	1	2	3	4	5
Number of straws	5	9	13		

2. How many straws are needed to make four houses in a row?
Write your answer in Lindsay's table.

3. How many straws are needed to make five houses in a row?
Write your answer in Lindsay's table.

Figure 4.25: From "Developing Problem Solving" 8-11 (Crust, 2005)

While it is not known that the PIM task Figure 4.26 was conceived as a variant of this genre – and it is aimed at slightly younger children – it does purport to assess “solving numerical problems¹⁴”. However, while the former requires students to spot and extend a number pattern, the latter simply requires them to count the sticks. There is an arguable speed or accuracy advantage if the sticks are counted in groups of 3 or 4, but this cannot be inferred

13 Indeed, one of the classes visited were engaged in an exemplary “plan a football tournament” problem solving activity.

14 See the teacher's book for the Progress In Mathematics 7 paper test - this task is Q4 on the paper (Clausen-May, Vappula, & Ruddock, 2004c)

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

from the answer – only one child out of 18 observed was clearly doing this. The question can be answered in reasonable time by simple, careful counting¹⁵. In fact, for pupils in this age range, solving the problem by multiplication would probably be less reliable than simple counting.

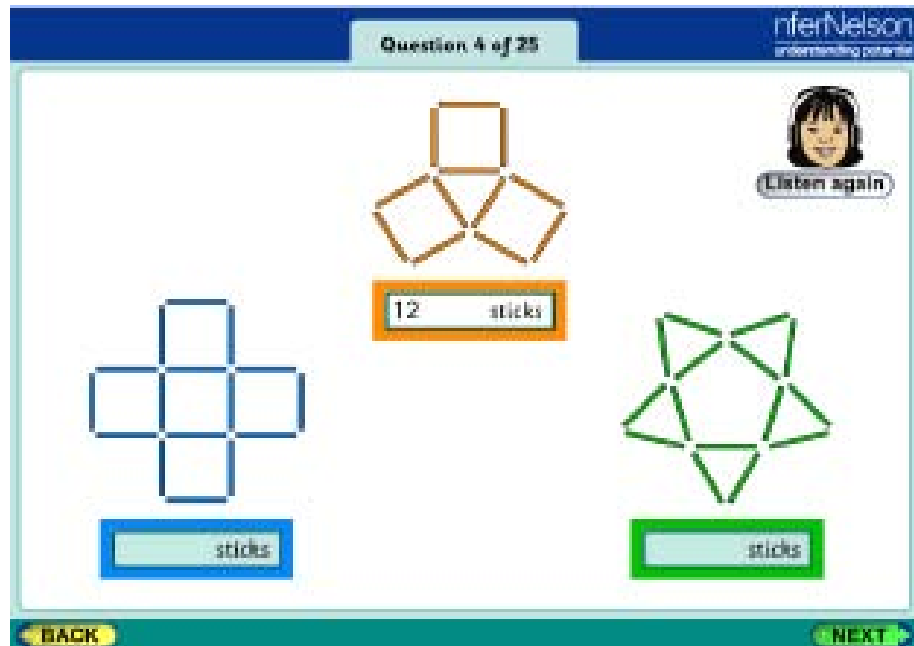


Figure 4.26: “Sticks” - from nferNelson “Progress in Mathematics” Age 7

If a task is to reliably assess a high level skill, it needs to be difficult or impossible to answer without using that skill. Figure 4.27 shows one example of a task that claims¹⁶ to test a high-level skill (using an efficient algorithm for subtracting 9). Obviously, pupils who can apply the rule will have a speed/accuracy advantage in this question – however, that assumes that they know how to subtract 10 easily. Many pupils were observed to subtract 10 by counting back on their fingers.

15 Of course, *Figure 4.25* could equally be solved by drawing the 5th house and counting – the full task is longer and goes on to ask students to formulate a rule. One constraint on testing problem solving and other high order skills is the need for self-contained 1-2 minute tasks.

16 From the teachers' guide: “(this task) assesses pupils' ability to follow the reasoning behind an algorithm for subtracting 9” (Clausen-May et al., 2004c)

Question 5 of 25

nferNelson
understanding potential

To subtract 9,
I subtract 10 and add 1 on.

Listen again

52 - 10 =

52 - 9 =

BACK NEXT

Figure 4.27: “Nine” - from nferNelson
“Progress In Mathematics” Age 7

Even when students were closely observed it was sometimes difficult to tell whether they were using the rule, whether their mistakes were faulty subtraction or if they were confused by the question. For example, an answer of “43” in the first box could mean that the student tried to subtract 10 by “counting back” and miscounted, or it could mean that they worked out $52-9$ correctly and entered it in the first box instead of the second – both cases were observed. The most common wrong response - “42,41” was sometimes the result of misapplying the rule by subtracting 10 and then *subtracting* one, but also occurred as a mistake when subtracting 9 by counting back. Some students were clearly confused as to what they were meant to do, sometimes correcting their answer when prompted to re-play the question. Others appeared to totally ignore the question and just answered both sums by “counting back”.

While a question like this may produce plausible psychometrics – students who are more confident and fluent in mathematics will, on average, do better – it does not reliably assess the higher order skill claimed, since a pupil ignoring the context and simply answering the two sums they see on the screen will get full marks – possibly with less scope for error than a pupil attempting to use the rule.

Some examples of design issues

Telling the time

The “clock” question, which appeared on both the digital and paper tests for Age 7, showed the greatest paper vs. digital discrepancy of any question, with the digital version facility level 35 percentage points below the paper version. A question on the Age 6 digital test, which also involved setting the time on a clock face, had no paper equivalent but also appeared extremely difficult, with a facility level of 23%. No immediate explanation for this was spotted during the design critique.

The task set to pupils was:

*The first clock shows the time when a train leaves London.
The train gets to Swansea three hours later.*

(On paper) On the second clock, draw the time when the train gets to Swansea

(On computer) At what time does the the train get to Swansea? Click on the arrow buttons to show your answer.¹⁷

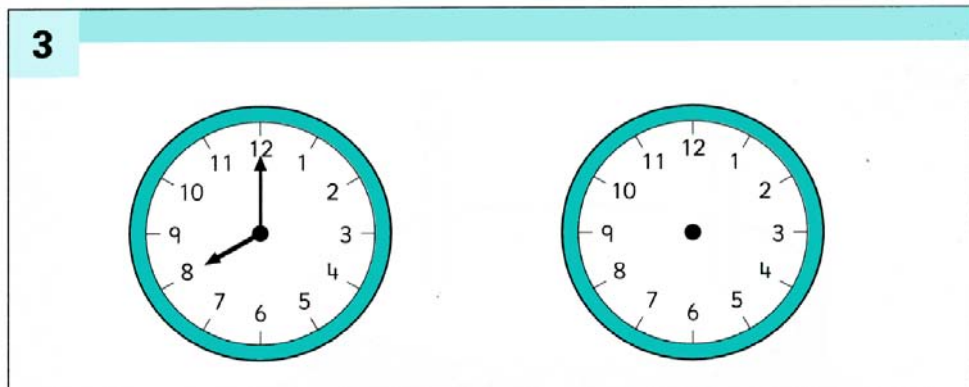


Figure 4.28: “Clock (paper)” - from nferNelson “Progress in Mathematics” Age 7. Students must draw hands on the second clock to show the time 3 hours after the time shown on the first clock.

In the paper version (Figure 4.28), the “second clock” was a plain clock face on which students could draw hands. On the digital versions (Figure 4.29), however, the “arrow buttons” adjusted the time shown on the clock in 15 minute jumps, with the hour hand accurately positioned to match the minute hand.

¹⁷ Although it may not have been the dominant issue in this task, this also illustrates how the language tended to change between the paper and computer tests.

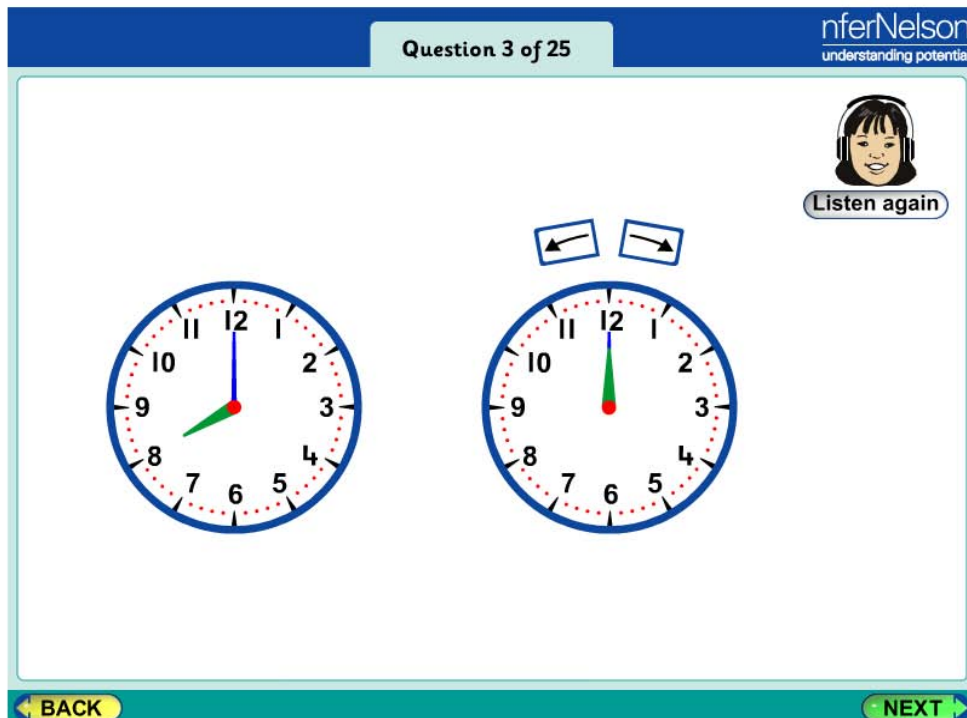


Figure 4.29: “Clock (digital)” - from nferNelson “Progress In Mathematics” Age 7. The computer version of the clock task. Here students click on the two arrow buttons to set the time on the second clock.

On the age 6 test, the digital version included a question using a similar clock face in which the task was simply:

Click on the arrow buttons to make the clock show half past three

The problem became evident when age 6 pupils were observed struggling to set their on-screen clock to half-past three: most, when asked, said that the hands should point to “6 and 3” which, while not precisely correct, would be a reasonable answer for a 6-year-old and acceptable according to the paper mark scheme. However, the following discussions (between pairs of pupils) were typical:

A: “Put the big hand to 6”

B: “Won’t let me”

or

A: “Go back”;

B: “No we need it [the big hand] to stay there – how do we do this?”

The problem was clearly that, at this age, pupils told the time using the mantra “the big hand points to... the little hand points to...” and were trying to set the time one hand at a time. The

4 - Analysis and Evaluation of Progress in Maths 6-14 Digital Assessments

inability to adjust the hands independently, and the “accurate” depiction of the hour hand's position were distracting students from answering the question.

In one school the observers were able to borrow a typical cardboard clock face and confirm that three pairs of students, all of whom had answered the question incorrectly, were able to set it to half-past three (or at least set the hands to 3 and 6).

Technically, it could be argued that this question is “fair”: at half-past three, the hour hand does **not** point directly to 6 and pupils should know this. However, the mark schemes for paper-based clock face questions at ages 7 and 8, for which the answers are all “on the hour” instructs markers to “allow each hand to be [not more than] halfway towards the next/previous number” so it seems unlikely that a question for 6 year olds would require higher precision. This stricture, therefore, appears to be an unintended artefact of the user interface design.

In the Age 7 question, students had to set the clock to 3 hours after 8:00, so the precise position of the hour hand was not such an obvious issue. However, when pupils were observed, at least 4 pairs who appeared to know the correct answer were either unable to enter it correctly or, while trying, lost concentration and switched to an incorrect answer. Pupils appeared to struggle to comprehend how the buttons set the time - one issue may have been that the hands moved in 15 minute jumps rather than “smoothly” so there were poor visual cues as to what each button did, especially when trying to move backwards.

The issue predicted by the design critique – that because the second clock started out at 12:00 rather than blank, some pupils would simply move it on 3 hours – was not evident, and the most common mistake in both the equating study and the observations was to simply set the clock to 8:00, suggesting that pupils had simply failed to listen to, or lost track of, the question.

This question had a relatively high language comprehension element, so the issues discussed previously regarding the efficacy of recorded voice prompts might be relevant – however, it was evident from the observations that user interface difficulties caused pupils to lose track of the task.

This task illustrates how the technical details of a user interface can disrupt pupils' engagement with the problem. Furthermore, this one task type could form the basis of a substantial study into how children learnt to tell the time and what the effect of various teaching aids was (some of the schools observed used traditional cardboard clocks, others used more sophisticated toys with realistically coupled hands or computer simulations). A complete maths test will contain many different user interfaces, each with potential unintended consequences.

Subtraction Sum

This was a straightforward question in which pupils had to complete a subtraction sum:

$$7 - 4 = \underline{\hspace{2cm}}$$

This appears to be the simplest sort of question to translate to computer – but in the equating study the facility level was 50% facility on computer vs. 65% on paper (and 69% in the National calibration tests on the paper).

Analysis of the equating study results show that the modal **wrong** answer was “11”, probably as a result of adding rather than subtracting – and that this mistake was twice as common on computer as on paper (22% vs. 11%). The answers “10” and “9”, probably indicating an erroneous attempt to add, were also more common on computer. The majority of equating study pupils gave a “plausible” (i.e. small integer) answer so entering the response did not appear to pose a problem.

The observations confirmed that 11 (adding, rather than subtracting) was the typical “wrong answer” and that pupils had no difficulty entering the answer.

The most plausible explanation here seems to be that pupils were simply more careless when taking the computer test – which might result from the lack of a teacher setting the pace by delivering the questions orally. Note that in this case, teachers are not supposed to read out the sum itself.

Square Shapes

In this question, the pupil's task was to identify, out of six possible choices shown in Fig. 4.26, the shape made from **8** small squares. The prompt was:

Sally put some squares together to make these six shapes. She made one of the shapes with eight squares. Click/Put a ring around the shape that Sally made with eight squares.

The equating study suggested that scores were significantly lower on the digital version (76% correct on digital vs. 94% on paper). All the pupils observed got the answer correct, so there was no observational evidence why this might be so.

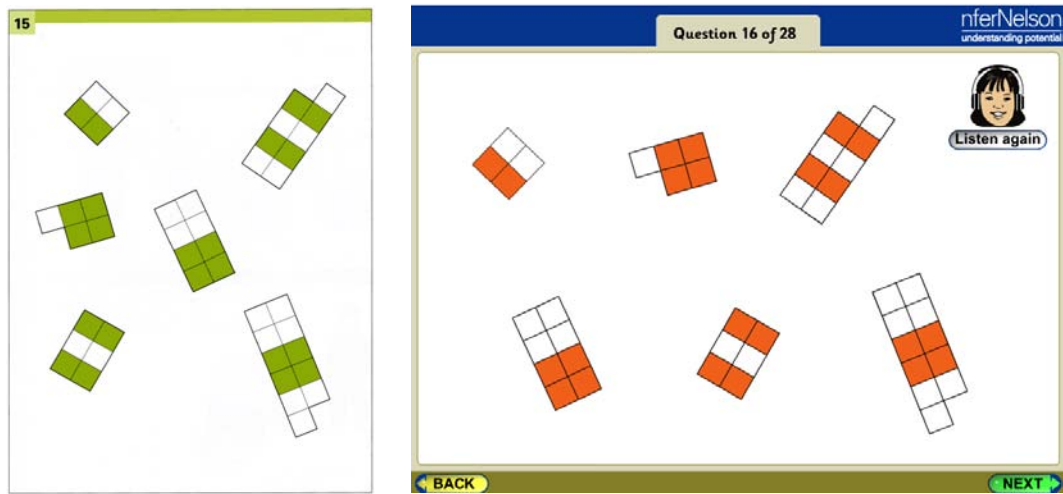


Figure 4.30: Paper (left) and digital versions of “square shapes” - from nferNelson “Progress in Maths” Age 7.

Possible explanations for this discrepancy could be:

- Pupils using the computer just made careless mistakes, or failed to listen to the question, due to the presentation of prompts. It is unclear why this should affect this question disproportionately, although the prompt is quite long in relation to the actual task (click on the shape with 8 squares).
- The design of the paper version places the correct shape in a prominent, central position, possibly to the benefit of pupils who guess the answer. The digital version uses a different layout.
- In the equating study, 9% of candidates failed to respond to this question on screen, compared to about 2% on paper. Possibly, some pupils failed to select any answer, due to the way the system handled multiple-choice tasks. Although no user interface problems were seen with this particular question during the observations, pupils did have trouble with one of the practice questions which used the same multiple choice technique: pupils clicked on their answer, and it was highlighted. However, they then expected some further response from the computer and, when nothing more happened, clicked on their answer again. This had the effect of de-selecting the answer. It is also possible that pupils carelessly clicked past the question without attempting it – something that would be far harder to do on the teacher-led paper test.

A similar effect was noted in a task requiring pupils to select the picture of a boat with the highest number written on its sail: as with *square shapes* the only obvious difference was an apparently minor change in the layout of the answers.

Other tasks highlighted by the equating study analysis

Opposite dice: Pupils were shown a die with five spots alongside a blank die representing the opposite side. They were given the information that the opposite sides of a die added up to 7 and asked to fill in the “missing dots”, by clicking a button to add dots to the blank die. The design critique noted that the context was quite complicated, but conversely that the user interface made it easy to “count on” from five to seven. In school, pupils were observed to start clicking instead of listening to the complete prompt, and hence failing to understand the question: in some cases, pupils started clicking immediately and were then confused when prompted to “fill in the missing dots” because, by then, there were no “missing dots”.

Subtract 45p: “Natasha has one pound. She buys a pencil for forty-five pence. Work out how much money Natasha has left.” No reason for the paper/digital discrepancy was observed. However, it was noted that the teacher's script for the paper version included the names of the questions, which did not appear on the digital version – and that in this case the name of the question, “subtract 45p”, gave a helpful clue as to how to solve the problem...

A new digital-only task

Figure 4.31 is an example of a new question type developed for the digital test. The question here was:

Look at this part of a number line. Move the arrow. Make it point to some different blue numbers. See how the number in the red box changes. What is added to the blue number to make the number in the red box? Click on your answer.

While there is no paper equivalent to compare this with, the facility level in the equating study was poor (40%). During the school observation, 14 out of 20 pairs were unable to answer correctly. The detailed observations suggested that the problems were due to the presentation and operation of the question – pupils simply did not understand what they were required to do.

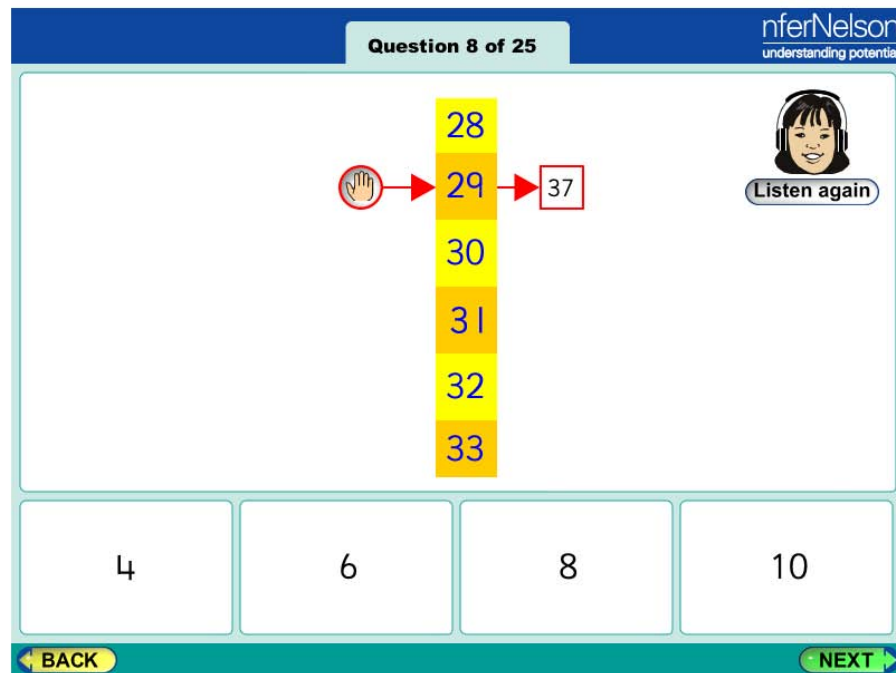


Figure 4.31: “Rules” - from nferNelson “Progress in Maths” Age 7 – digital version

One common problem was that pupils misidentified the “red box” and “blue numbers” referred to in the question. Questioning by the observer confirmed that at least 7 of the 20 pairs had problems with this¹⁸ - typically thinking that the orange squares on the number line were the “red box” and that the four responses along the bottom were the “blue numbers”. When viewed on a typical TFT display, especially at a slight angle, the source of this confusion was clear – it was far from obvious which numbers were blue, and the “red box” was far less prominent than the chequered number line.

The design of this task raises questions about the motivation for adding an interactive element to such a straightforward task. The information revealed by moving the slider up and down is entirely irrelevant to the task, as the prompt confirms that the “rule” is a simple addition, so only one pair of numbers is needed to answer. The scenario remains completely abstract: the interaction does not place the task in any sort of valid context, it doesn't even provide any clear visualisation to suggest that a function is being applied to an input number. While there may be value in embedding a task in a context which adds some cognitive load in addition to the underlying mathematics, in this case the “context” serves purely to obfuscate the problem, and necessitates a long, complex prompt.

¹⁸ Colour notes: in Fig. 4.31, the vertical “number line” has alternate yellow and orange squares and contains the “blue numbers”. The square around “37” is the red box. The 4 numbers along the bottom are the possible answers.

4.6: Schools and equipment provision

Not enough schools were visited to draw significant inferences about ICT provision, but informal notes were taken about the equipment provided and environment in which the observations took place.

One issue was a general lack of whole-class computing facilities for younger pupils: of the six schools visited, only one had a “computer lab” equipped with desks and chairs suitable for 6-7 year-olds. Computing facilities for this age group typically consisted of a small number of computers in each classroom. These were clearly used regularly and effectively, in some cases pupils were free to use them at any reasonable time, but would not be suitable for a whole class wishing to take an online test. Some schools hoped to obtain a “class set” of laptops, which would help address this need. Dedicated computer labs were more common in “junior” schools catering for a larger age range and tended to be equipped for older children – in one case, with high desks and stools which would have been unsuitable for use by 6 year-olds.

Even where desks and chairs were of suitable height, computing equipment was, universally, generic equipment designed for adults. This seemed generally acceptable with the possible exception of mice. When an adult is using a mouse, and reaches the edge of the desk or the extent of their reach, they will typically pick up the mouse in their palm and re-position it: an experienced user will probably be able to do this even while holding down a button to drag an icon. This manoeuvre is nearly impossible for a young child with small hands using a full-sized adult mouse. While the pupils observed worked around this obstacle without complaint, it is a potential source of distraction and extra “cognitive load”, so the suitability of full-sized adult mice for young children would bear further investigation.

Maintenance of peripherals, particularly mice, is essential – in one case a faulty mouse was generating spurious double-clicks and causing the pupil to skip over entire questions; in another case an entire computer lab was equipped with mouse mats featuring large areas of plain colour on which the accompanying optical mice simply didn't work.

Provision of headphones was also an issue for these tests, which relied on spoken prompts. Some schools provided bulky, old-fashioned “cans” which appeared somewhat ungainly for small children, while others provided modern, lightweight headphones whose foam rubber ear pads had long since vanished.

4.7: Weaknesses of the experimental model

The work described here was largely based on existing data from nferNelson's own equating study and was further constrained by the need to complete the school observations during the summer term of 2006.

In particular, as noted above, the pupils taking part in the equating study were given **both** digital and paper versions of the test with no attempt to control the order or timing of the two tests – with the result that some pupils took both tests on the same day, others took them several weeks apart. Had the data come from a larger number of schools, these effects could possibly have been investigated and compensated for, but with only 9 schools involved, and the order and timing of the two tests usually applying to whole schools, it was impossible to distinguish between order of testing and other possible school-wide effects, such as faulty equipment or generally poor IT skills.

Given complete freedom, more resources and a measure of hindsight, a better methodology might have been as follows:

1. Conduct the design critique of the computer tasks to identify potential issues
2. Conduct the small-scale school observations – ideally, this should be done independently rather than by the authors of the design critique
3. Address any clearly identified design faults in the questions arising from the design critique and small-scale trials. If the changes are major, another round of small-scale trials might be needed to verify that the changes were effective
4. Trial the digital tests alone with a sample of 100-200 students who had not previously taken the test. Since, in the case of PIM, there is a large, existing set of data on the paper tests (over 2000 students per test from the original calibration exercise) it might not be necessary to re-trial the paper tests – the digital results can simply be compared with the existing data. Alternatively, a cross-over study in which each student took half of the questions on paper and the other half on computer could be used – however, this would require involving sufficient schools to ensure an adequate “cross-over” sample should some schools fail to take both parts. Control, or at least record, the order and time interval between tests
5. Ensure that any technical problems such as poor internet connections, faulty equipment or software bugs are noted, especially where these issues affect an entire school

4.8: Conclusions

Implications for our research questions

Research question B (Section 1.2) asked “What are the effects of transforming an existing paper-based test to computer?” and a key objective of the *Progress in Maths* study was to ascertain whether the difficulty and mathematical validity of the new computer-based test was “equivalent” to the existing, traditional assessment. The practical question was whether a relatively small “equating study” could be used to apply the calibration data for the original test – originating from a far larger sample – to the new computer-based test.

While this study failed to find unambiguous evidence for a systematic effect on difficulty, there were a number of cases where design decisions arising from the change of medium had an effect that was evident from the observations, the equating study data or both. One major change – the move from teacher-read prompts to recorded voices – seemed particularly likely to have affected performance across the test.

Was it reasonable to expect the tests to be equivalent? Considering our final research question (D) what design and development considerations could have helped attain equivalence?

Expectations of equivalence

It is always difficult to make a clear distinction between a student’s mathematical skills and other factors, such as vocabulary, reading ability, “listening ability”, dexterity and attention span. At ages 6 and 7 the range of ability encountered in a “typical” class is wide: the more able students observed in this study were proficient readers who could probably have coped with entirely written questions, whereas the lower end of the normal ability range (not necessarily considered to have special educational needs) could not read and appeared to find comprehending the verbal prompts demanding.

In such an environment, **any** change in the presentation mode of a test runs the risk of changing the psychometrics of the question. Having fixed, pre-recorded questions instead of questions read out by a familiar teacher, with liberty to repeat or paraphrase questions is such a fundamental change that it is unlikely that the tests will ever be completely equivalent.

The question is, therefore, is the disparity predictable; can it be compensated for and is such a correction valid?

The data from the equating study suggested an overall reduction in performance on the digital test but, mainly due to concerns over the validity of the data, it was not possible to either verify the significance or quantify this trend. The overall impression from the observations was that the change in presentation mode made the digital test more prone to random,

careless mistakes than the teacher-paced paper test. If this could be quantified by a more rigorous trial, it might be reasonable to apply a correction to the score.

What was clear, however, is that there are additional effects which are highly dependent on the design of individual questions – even ones which appear almost identical to their paper versions. The most striking example was the *clock* task where the observational evidence clearly revealed how the apparently simple user interface had complicated the task. This could make a global “correction” risky. Unless paper and digital tests were constructed using only those items proven to either have the same difficulty or to fit the “global” correction model, the safest approach would be to calibrate digital tests separately to the paper versions, and not attempt to present them as the “same” test.

Design guidelines for equivalence

Here are some questions that might help determine, at the design stage, whether a digital item might not be equivalent to the paper original:

- Have the illustrations been changed – do the new illustrations show any numbers, quantities or mathematical artefacts that might cause a distraction?
- Is the question prompt longer? Is the language still appropriate to the age group?
- In the particular case of the PIM tasks with audio prompts:
 - Are there any words in the prompt that a teacher might legitimately replace with “easier” ones?
 - Are there any pauses in the prompt that a teacher, watching the reaction of the class, would be able to judge better?
- Where multiple answers are shown, does the digital version present them in the same order/layout? Does the layout in either version “draw the eye” to a particular option?
- Does the digital version impose any additional sequencing on the answer (e.g. revealing the possible answers one at a time) or provide a default answer not present on paper (e.g. the clock pre-set to 12:00 rather than a blank face)?
- Does the digital version impose new constraints on the answer – such as the clock question which forced the position of the hour hand to be accurately linked to that of the minute hand?
- What are the wrong answers that might be expected – and can the student enter those? (E.g. in the “scales” practice task several students incorrectly read the scales as $4\frac{1}{2}$ instead of 5 – but were then unable to enter that answer)

- Does the student now have to interact with the question to obtain required information which was presented “up front” on paper?

The importance of close observations

By observing small groups of children interacting, in pairs, with the questions, this study was able to obtain detailed feedback on possible problems with the design and implementation of the questions. In a few cases (such as the *clock* task), this process made it clear that pupils were failing to engage with the intended mathematical content of the task because of identifiable flaws in the digital implementation: something which would have been difficult to infer by simply analysing responses in bulk.

This observation step is an essential part of the process of developing a piece of software. The aim is to ensure that pupils' can interact successfully with the test questions and to identify any areas of the software design which require refinement. Ideally, of course, the same process should be followed for a conventional paper test, but where software is involved the potential for introducing unintended complications is far more significant, since the design is more complex and there is often an additional step in which the intentions of the task designer are interpreted by a programmer.

Having pupils work in pairs is useful, since it encourages them to vocalise their thoughts without constant interruption by the observer. Allowing the observer to intervene is also important – it allows the ad-hoc plausibility testing of hypotheses as well as prompting to determine the root of the problem (if pupils are shown how to interact with the computer, can they engage with the question? If they are given the correct answer, can they input it?) Both of these, along with the small numbers required to enable close observation, prevent the collection of reliable psychometric data on the performance of the questions, so it is important that this step is separated from the test calibration process. The development cycle must include sufficient time for this small-scale qualitative trialling to be completed before embarking on quantitative trials.

It would be valuable experience for programmers and designers developing such tests to be involved in such observations. There is a new skills-base to be established here: interactive test designers whose expertise covers both mathematical education and interactive software design. The author's experience is that these undertakings often reveal a divide between pedagogical experts with little or no software design experience and programmer/designers with no pedagogical background, working in an “author/publisher” relationship.

5: Computerising mathematics assessment at GCSE: the challenge

On two occasions I have been asked,—“Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?”

I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.

Charles Babbage (Babbage & Campbell-Kelly, 1994, p. 67)

5.1: Introduction

One of the motivations behind the studies leading to this thesis was the enthusiastic plans by QCA to move towards computer-based testing at GCSE, on a timescale which would appear to rule out any radical re-design of the existing GCSE curriculum. So it seems reasonable to assume that a near-future GCSE will rely heavily on tried and tested styles of task, even if some new task types are introduced.

In Chapter 4, it became clear that even fairly simple tasks, with single numerical or multiple choice answers, raise issues when converted to computer, posing the question of how the more sophisticated tasks currently seen at GCSE might be affected. To this end, it is worth looking in some detail at the range of task types currently seen on GCSE papers and considering how these might be computerised.

Also, before trying to devise elaborate methods of faithfully recreating particular features of GCSE on computer, this is an opportunity to critique the design of existing questions and consider how effective these features are, and whether they might contribute to the assessment aspirations discussed in Chapter 2.

Note: out of necessity, the work of this chapter was conducted at an early stage in this project, before embarking on the experiment described in the following chapter. The state of post-14 mathematics in England is extensively reviewed in the Smith Report (2004) and one evaluation of the piloting of the new assessments suggested therein can be found on the *EMP project* website (Murphy & Noyes, 2010)

5.2: The current state of GCSE mathematics

Scope of this work

The limited analysis of GCSE mathematics performed for this work is intended to identify the range of presentation/response types, the allocation of marks in a “typical” GCSE paper, and a small set of typical task types suitable for experimental computer-based delivery. It is also informed by the author's involvement in the debate on the proposed introduction of assessment of “functional mathematics” and how that contrasts with existing practice. Experienced GCSE markers and examiners were consulted on issues such as the interpretation of mark schemes.

An in-depth analysis and critique of GCSE mathematics, across different awarding bodies, ability ranges and syllabus specifications is beyond the scope of this chapter – and would not be timely as the system is currently in an unusual state of flux¹⁹. The summary below is intended as a brief overview for those unfamiliar with GCSE in general or mathematics in particular.

In order to obtain an impression of the types of responses and marking strategies that a computer-based system might need, several specimen AQA GCSE papers were examined, and one was analysed in detail. Later, a similar analysis was performed on a pair of “live” AQA GCSE papers – the figures quoted below come from the latter analysis. Since these papers are rigorously standardised it is reasonable to assume that this is representative. These figures are intended to be indicative of the number of questions or the proportion of total marks connected with particular styles of response and mark scheme. Since, as mentioned below, the GCSE mark schemes are quite complex, these figures include an element of judgement and uncertainty.

A typical GCSE mathematics examination

The General Certificate of Secondary Education is the main qualification taken at the end of compulsory schooling (age 16) in England and Wales. Each subject is examined separately

¹⁹ For instance, the coursework option has recently been abolished, the system is being reduced from three to two tiers, a new National Curriculum program of study has been introduced and “functional mathematics” is being incorporated into GCSE.

5 - Computerising mathematics assessment at GCSE: the challenge

and awarded a grade from A* to G. There is no formal concept of “graduation” or an overall cross-subject grade for individual students, although “Five or more passes at grade C or above, including English and Mathematics” is a common rule of thumb for a “successful” student. Various systems for combining results across subjects into a single score have been used for school accountability purposes.

Examinations are set and administered by five independent Awarding Bodies (examination boards) within strict standards set and policed by the Qualifications and Curriculum Authority (QCA)²⁰.

Mathematics

Most of the examples are taken from the 2003 AQA mathematics syllabus - this is a typical “traditional” syllabus that largely depends on the final examination. Most boards also offer a modular alternative. The GCSE specifications are tightly controlled by the QCA and are fairly consistent between examination boards.

The examination consists of two 2-hour papers with the use of calculators permitted only on the second paper. Both papers cover the entire syllabus – although, predictably, the non-calculator paper has a bias towards testing arithmetic and computational skills. Together, the two papers aim to cover all the major topics on the syllabus rather than “sampling” selected topics. There is an element of sampling, but at a detailed sub-topic level. So, for example, there will always be a question on interpreting data in chart form, the likely variation between tests being whether the chart is (say) a pie chart, bar chart or stem-and-leaf plot.

In order to cater for a wide range of abilities, papers were (at the time of writing) “tiered” into Foundation, Intermediate and Higher tiers, with entrants at the lower tier ineligible for higher grades. The Foundation tier has more emphasis on basic number skills, the Higher tier more algebra and trigonometry. The papers are “ramped” (i.e. the difficulty increases gradually throughout) at least insofar as the more advanced topics appear later in the paper. Roughly speaking, the Intermediate tier paper corresponds to the last half of the Foundation paper and the first half of the Higher paper.

This study concentrates mainly on the intermediate tier paper.

Each paper consists of 20-25 questions with an average 4-5 marks per question (the total marks are usually arranged to be 100). Typically each question is in 1-3 parts each worth 1-3 marks (the modal value is 2 marks per question part).

Some examples of GCSE questions can be found later in this chapter, and the next.

²⁰ This arrangement was in flux at the time of writing: a new body “Ofqual” is responsible for regulating examinations while QCA (now QCDA) retains responsibility for curriculum development and specification.

Responses and marking

Responses

The most common response format is a short written answer – often consisting of a single number, or, less frequently 4-6 lines of text. A few questions will ask the student to draw or complete a graph or chart (usually on supplied axes) and there is usually at least one “construct with ruler/compass/protractor” question.

Figure 5.1 shows the relative frequencies of answer types in the pair of papers analysed in detail. “Subtask” is used here to mean any question, or part of a question, which requires a response.

	Response format	Frequency
Written	Answer line - no units	41
	Answer line - units given	36
	Writing space - lined	12
	Ordered pair	2
	Missing words/numbers	2
	Complete the table	1
Drawn	Axes - fixed scale	3
	Complete the diagram	2
	White space (for drawing)	2
Total (subtasks):		101

Figure 5.1: Relative frequency of answer formats

90% of the sub-tasks on the papers analysed either had lined space for working (in addition to the answer) or allowed a multi-line written answer²¹. Students are given a general instruction to “always show working” at the start of the test - while 16% of sub-tasks specifically requested students to show working or otherwise support their answer.

Mark schemes

The mark schemes are quite complex – the AQA schemes studied involve concepts such as accuracy marks, method marks, dependent marks, bonus marks, independent marks and “follow through” (where allowance is made for correct work following on from an incorrect result). There are numerous “special cases”, alternative answers and multiple criteria for part marks. Using these schemes requires both familiarity with the conventions and experience of mathematics and teaching. Marking is performed by professional markers, often teachers, employed and monitored by the boards.

²¹ The exceptions were usually questions involving graphs, diagrams or fill-in tables

The mark schemes for most sub-tasks (68%) allow partial credit based on “working that could lead to a correct answer” or other marks that can be awarded without a correct answer to the subtask. However, unless the question *specifically* reminds candidates to show their working, these marks are usually awarded by default if the final answer is correct.

5.3: Weaknesses of the GCSE format

Fragmentation

Typical GCSE tasks are either short, or composed of short sub-tasks – with no substantial chains of reasoning. Sub-tasks may share a context, topic or a resource (such as a diagram or table of data) but they can usually be answered completely independently of the other parts - such as parts (a) and (b) in *Figure 5.2*. Consequently, students are being tested on discrete items of technique and knowledge rather than their ability to combine several such skills to solve a substantial problem.

One indication of fragmentation is the number of marks available per sub-task, since the available marks usually correspond to identifiable steps in the solution. Also, since designers usually intend the available marks to be proportionate to the time required, this can also provide a coarse estimate of the *unsupported reasoning length* – the length of time students are expected to spend working on a problem without further prompts and sub-questions.

The distribution of marks per sub-task in the two papers analysed is shown in *Figure 5.3* - showing that almost 80% of sub-tasks are worth 1-2 marks. It is not possible to accurately determine the reasoning length for each question without actually observing a sample of students. However, a coarse estimate can be made by assuming that the time taken to complete each question is roughly proportional to the marks available for it. On a 2 hour test, this would give a modal value of 1.2 minutes per sub-task.

This may be symptomatic of the way the syllabus is regulated to the point of specifying how many marks must be allocated to each specific item of content knowledge. It represents the “safest” path for test design in terms of consistency and defensibility of results, but does not assess pupils' ability to sustain substantial chains of reasoning.

Topic area/activity	Primary focus		Total dependent	
	Marks	% of Total	Marks	% of total
Arithmetic	25	12.5%	84	42.0%
Arithmetic (calculator)	9	4.5%	12	6.0%
Accuracy (formalisms)	3	1.5%	5	2.5%
Accuracy (other)	1	0.5%	1	0.5%
Identify mathematical relationships	17	8.5%	20	10.0%
Measurement/drawing	4	2.0%	4	2.0%
Explain + justify	13	6.5%	15	7.5%
Deduction	5	2.5%	11	5.5%
Quantitative awareness	10	5.0%	10	5.0%
Spatial reasoning	9	4.5%	10	5.0%
Apply previously deduced rule	2	1.0%	2	1.0%
Manipulate expression (algebra or other)	30	15.0%	31	15.5%
Apply supplied formula	5	2.5%	5	2.5%
Apply standard formula	10	5.0%	10	5.0%
Formulate	4	2.0%	5	2.5%
Understand representation	9	4.5%	23	11.5%
Choose representation	1	0.5%	1	0.5%
Other technical knowledge	43	21.5%	64	32.0%
Total marks	200			

For each mark on the papers, a single primary topic/activity focus was identified. The "Total dependent" column includes other marks judged to have some secondary demand for that activity.

Figure 5.4: Distribution of marks by topic/activity in the GCSE sample

Technical vs. strategic skills

The ability to select the most appropriate techniques to solve a problem, to choose the best representations to use, and with which to communicate the result, are key aspects of mathematical performance. As a consequence of the fragmentation noted above, the correct mathematics is normally implied by the question. Representations are usually pre-determined, not chosen by the student, with instructions such as "complete this table" or "design a tally chart to show the above information". Thus, these key strategic skills are rarely assessed by GCSE tests.

Figure 5.4 gives an impression of the distribution of marks amongst topic and activity types on the papers analysed. For each mark, the question and mark scheme were examined to identify the primary mathematical activity involved (one per mark) and any other secondary activities needed to attain that mark (so, unsurprisingly, many marks require some element of mathematical knowledge plus correct arithmetic). The topic headings were chosen based on the range seen in the papers, rather than trying to force the questions into a framework for which they are not designed. It can be seen that, while there are some topics that might fall under the heading of "strategic skills" the marks are dominated by arithmetic, manipulation and technical knowledge.

Mathematics for an IT-driven world

Much time is devoted to doing calculations, drawing graphs, completing tables, remembering definitions and even constructing drawings with ruler and compasses - but rarely on interpreting or explaining the result. For example, the typical “statistics” question involves drawing a chart or table and *ends* with calculating the mean, median or mode – without any discussion of its meaning or interpretation in the context of the question.

The role of the clerk or “human computer” who can perform routine calculations without regard to their significance or context has long been obsolete. Technology has made a wide range of powerful tools and techniques available to everyone. In the world of educational research, for instance, it is no longer expected that every researcher should be able to manually calculate, or write their own software for, statistical tests. The key skill is to know what techniques are available, what their applications and limitations are, to be able to interpret the results and to be able to spot implausible results when things go wrong.

However, at GCSE, the focus is still on being able to perform a wide range of techniques manually (or with minimal help from a calculator). There is the occasional task on the more sophisticated use of the calculator, such as:

(i) Use your calculator to find $\sqrt{28.9^2 - 9.24^2}$

Give all the figures in your calculator display

.....

(ii) Write your answer to 3 significant figures

.....

AQA Spec. A Intermediate Paper 2 – June 2003 (AQA, 2003a)

This is typical of such tasks. The two key concepts are dealing with the correct order of operations and rounding the result to a given number of significant figures (though rarely relating this to the precision of the supplied input numbers).

Order-of-magnitude estimation and plausibility-checking are essential skills when using IT in mathematics, as part of the eternal vigilance against bugs and “garbage in/garbage out” and to enable critical understanding of figures cited in scientific and political debates. The estimation tasks at GCSE are usually limited to “approximate arithmetic evaluation” of pre-defined expressions in which each operand can obviously be rounded to a single significant figure. The following are typical:

Find an approximate value of $\frac{2897}{21 \times 49}$

You **must** show all your working

AQA Spec. A Intermediate Paper 1 – November 2003 (AQA, 2003b)

Use approximations to estimate the value of: $\sqrt{\frac{9.98}{0.203}}$

You **must** show your working

AQA Spec. A Intermediate Paper 1 – June 2006 (AQA, 2006)

Note that this is one of the rare cases where the mark scheme penalises lack of method²²: an answer of “3” to the first example would score nothing.

These questions represent just one type of estimation task (approximate arithmetic). Other forms of estimation (Johnson, 1979), including awareness of orders of magnitude, judging the reasonableness of results and “Fermi estimates” such as:

“Jane says that there are about a million primary schools in England – is she right?”

are absent from GCSE.

Plausibility of “realistic” concepts

Questions are often set “in context” rather than stated in bald mathematical language. For example, variations on the following regularly appear:

“Mrs Jones decides to distribute her inheritance of £12 000 amongst her 3 children in the ratio 7:8:9”

AQA Spec. A, Intermediate, Paper 2, June 2003 (AQA, 2003a)

However, it is usually clear that the context has been contrived to fit a particular statement in the syllabus and is not an authentic situation in the real world to which mathematics would be applied²³. The fact that 7+8+9 is a multiple of the number of children emphasises the

²² These were taken from a paper which does **not** allow the use of calculators

²³ The quotation at the start of chapter 2 would suggest that this issue pre-dates GCSE by a few millennia.

5 - Computerising mathematics assessment at GCSE: the challenge

artificiality of the situation (or perhaps Mrs Jones is a good functional mathematician and likes to keep things simple). The fact that her fortune is so easily divisible by $7+8+9$ is also unrealistic (calculators were allowed on this paper, and the mark scheme for this question did not reward students who used mental arithmetic). The same mathematics could be assessed as part of a more plausible “functional mathematics” question: given the ages and financial situations of her children, how should Mrs Jones choose to divide her money? How should she describe this in her will so it does not depend on the final value of the inheritance? Alternatively, a more authentic context for this mathematics might have involved, for example, mixing concrete, where sand, gravel and cement do have to be mixed to a given ratio, and the quantities of each needed for a given total amount are required.

Other questions on the GCSE papers showed a disconnect from the real world. For example:

The table shows the exchange rates between different currencies:

£1 (Pound)	is worth	1.64 euros
\$1 (Dollar)	is worth	1.05 euros

- Jane changes £400 into euros. How many euros does she receive?
- Sonia changes £672 euros into dollars. How many dollars does she receive?

AQA Spec. A Intermediate Paper 2 – June 2003 (AQA, 2003a)

This is reasonable as an exercise involving rate conversions, but the question has been designed from a pure mathematical perspective which sees all rate conversion questions as equivalent, whether they involve quantities, distances, times or currencies. In reality, although the statement “£1 is worth €1.64” might appear in the press²⁴ this does not mean that Jane can get 656 euros for her £400 at a *bureau de change*. The real task facing Jane at the airport will be:

Den's Currency Exchange		
Currency	We Buy	We Sell
\$ US Dollar	£ 0.533	£ 0.592
€ Euro	£ 0.659	£ 0.731
No commission!		

- How many Euros (€) would you get for £ 500?
- How many Pounds (£) can you get for \$ 700?
- How much would you have to pay, in Pounds and Pence, to get exactly € 550?

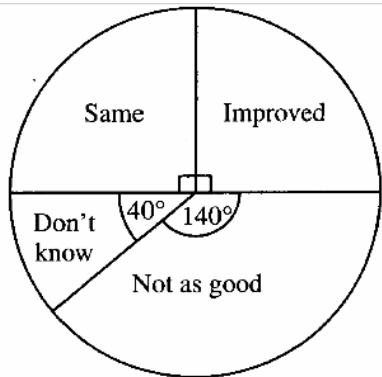
²⁴ ...although, at the time of editing this chapter, this seems mere nostalgia...

5 - Computerising mathematics assessment at GCSE: the challenge

Here, the central mathematical concept is the same, but several other steps are needed: knowing or deducing which of the two rates to use in each case; dealing with prices to three decimal places and deciding the correct calculation to use for part (c). The GCSE question is likely to be easier – but is inadequate preparation for the real problem.

Representations are often used in contrived ways: for example, Figure 5.5 shows part of a question which requires candidates to accurately extract “raw data” from a pie chart. This clearly tests the basic concept of a pie chart, and several other mathematical skills, but uses the pie chart as an inefficient (even when the angles have been conveniently written on) way of communicating numerical data, rather than its authentic role as an aid to visualising data. Note that, even in the complete question, the candidate is not asked to draw any conclusions from the data.

180 other people are asked the same question.
The results of this survey are shown on the pie chart.



(i) How many people answered “Improved”?

.....
.....
.....

(2 marks)

(ii) What is the probability that a person picked at random from this second survey answered “Don’t know”?

Give your answer as a fraction in its simplest form.

Figure 5.5: Extract from AQA 2003 specimen GCSE papers

The “context” problems seen on GCSE papers are probably more engaging than tasks presented in bare mathematical language, and likely to be more difficult as a result. However, they rarely assess problem solving skills in realistic contexts in the ways discussed in Chapter 2. The designers have introduced contexts and characters to the question, but do not have appeared to consider the real-world plausibility of the task.

Mathematical validity

In the examples above, the mathematical principles are perfectly valid: the criticism arises from the role, plausibility and practicality of the context and the form of the question. In some cases, however, it appears that the question could introduce or reinforce mathematical misconceptions.

Firstly, the focus on assessing *simple uses of quite sophisticated mathematics* (rather than the *sophisticated uses of relatively elementary mathematical tools* (Steen & Forman, 2000) advocated by some proponents of functional mathematics) can result in the routine use of “special cases” to produce accessible tasks. Does regular exposure to such tasks leave students unable to deal with more general cases? Do they recognise why the cases they have seen are special?

Trivial examples include the tendency for calculations to feature conveniently round numbers, thus avoiding issues of appropriate accuracy which could be critical to any real situation. This was noted in 5.3 above and is also noticeable in Figure 5.5 (note how the number of people represented by the pie chart is exactly 180 so 1 person = 2 degrees). Another example is that questions on medians and quartiles typically feature a convenient number of cases ($4N+3$) such that there is always an actual data point corresponding to each quartile. This introduces a potential misconception (there always has to be a data point at each quartile) and does not provide evidence that a pupil could handle the calculation in the general case. This is not to suggest that all questions should feature gratuitously awkward numbers (and it is not the case that all questions on the calculator paper examined feature trivial arithmetic) but the examples above reduce the question to a special case and could possibly foster misconceptions.

Some questions appear to embody more serious misconceptions or fallacies, and invite students to apply mathematics in fundamentally invalid ways. For example one sample GCSE statistics question (Figure 5.6) invites students to “prove” a hypothesis by noting a difference between two relative frequencies without considering whether the results were significant. Figure 5.7 Shows a Chi-squared test on the data, using a free online tool found on the web: although this is not part of the GCSE curriculum, this is the sort of tool which could be made available on a computer-based test. Pupils could be aware of the existence and importance of tests of statistical significance without being expected to learn the mechanics of performing them²⁵.

25 Even researchers rely on a table or pre-written software for the step of converting Chi-squared to a probability.

- (ii) Seema records information from a sample of 30 boys and 20 girls. She finds that 13 boys and 12 girls eat healthy food. Based on this sample, is the hypothesis correct? Explain your answer.

.....

.....

.....

.....

.....

(2 marks)

Figure 5.6: Question from AQA Mathematics Specification A Paper 2, November 2003. Are girls more likely to eat healthy food than boys? (Yes, according to the mark scheme)

	Yes	No	Total
Boy	13	17	30
Girl	12	8	20
Total	25	25	50

Degrees of freedom: 1

Chi-square = 1.33333333333333

For significance at the .05 level, chi-square should be greater than or equal to 3.84.

The distribution is not significant. *p* is less than or equal to 1.

Figure 5.7: Chi-squared test on data from Figure 5.6 using a free web-based tool

5 - Computerising mathematics assessment at GCSE: the challenge

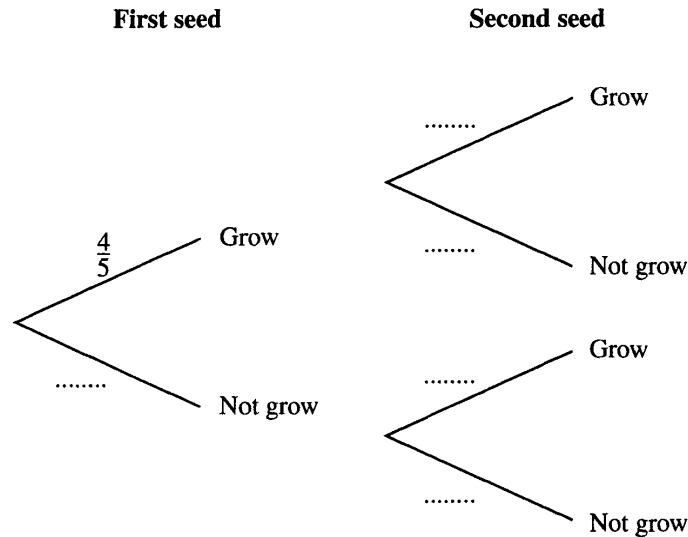
Another common generic question type requires students to show that they can calculate the probability of two events happening together by multiplying the probabilities, a rule which makes the crucial assumption that the two events are independent. This basic task regularly appears in various contexts – such as throwing a pair of dice, tossing coins, or throwing darts – in which this assumption is reasonable. On one occasion, however, the same underlying task was used in the context of a number of plant pots each containing two seeds (Figure 5.8). Given the probability of one seed germinating, students were asked to predict the number of pots in which neither seed would grow. The question included a “tree diagram” to be filled in, so it was clear that the probabilities were to be combined in the usual way. Since there are a number of obvious factors (water, temperature, sunlight, soil composition...) which might affect the germination of **both** seeds in a particular pot it is clear that, in this context, the events cannot be assumed to be independent, and the rules of combining probabilities for independent events are invalid in this context.

These examples were taken from a small review of a limited number of papers from one board, the primary aim of which was to inform the design of the trial online testing system. A more in-depth study would be required to determine whether they represented a serious, systematic problem with the validity of GCSE mathematics. However, these examples illustrate some of the issues which can arise from the practice of taking a pure mathematical exercise and constructing a largely cosmetic context around it without considering the combined validity of the mathematics and context.

A primary school teacher plants some sunflower seeds.
 She plants two seeds in each pot.
 The probability that a seed grows is $\frac{4}{5}$

The probability tree diagram shows the outcomes for the two seeds in a pot.

(a) Complete the probability tree diagram.



(2 marks)

(b) (i) What is the probability that both seeds grow?

.....

(2 marks)

(ii) What is the probability that at least one seed grows?

.....

(2 marks)

(c) There are 25 children in the class.
 Each child takes home one pot planted with two seeds.
 The teacher plants some extra pots in case **neither** seed grows in some of the children's pots.
 How many extra pots should the teacher plant?
 Your answer should refer to probabilities you have calculated.

.....

(2 marks)

Figure 5.8: Probability question from an AQA Specification A GCSE Paper

Allocation of marks

As noted above, the assessment of reasoning and working is a significant element of GCSE mark schemes, which include details and examples of the types of calculation that markers should look for.

From the eAssessment perspective, there are two practical concerns surrounding such “method marks”:

- the need to analyse working and assign partial credit greatly complicates automatic marking. Not only do algorithms – probably heuristic rather than analytic – have to be devised for each question type, but the issue of “follow through” (where credit is given for a correct calculation based on an incorrect result from a previous step) has to be addressed. Marking software cannot, therefore, assume a one-to-one mapping between question, response and mark²⁶
- working may be difficult to express in plain text, so tools will have to be provided for other response formats. The process by which the student inputs their “method” may either increase the “cognitive load” of the question – usually making the question harder – or alternatively make the question easier by providing on-screen forms or templates that suggest the correct method (such as separate boxes for the numerator and denominator of a fraction).

The tendency at GCSE is for questions to be broken down into sub-tasks, each with a separate prompt, space for working and space for a response. A typical mark scheme for such sub-tasks will allocate one or two “method” marks for sight of the correct method and one “accuracy” mark for the correct final answer, but if the final answer is correct then the corresponding method marks are usually awarded by default. Although the front of the test booklet advises students “in all calculations, show clearly how you worked out your answer”, students who produce the correct final answer are only penalised for missing working or incorrect method if:

- the individual question specifically asks for method, working or reasoning *and* this is reflected in the mark scheme
- or**
- the student's work shows that the “correct” answer was obtained from an incorrect method – if there is any ambiguity students are to be given the benefit of the doubt.

²⁶ There are also a few cases where a particular mistake, such as incorrect notation or the omission of units, is only penalised once across the whole paper, which also breaks the strict response-to-mark correspondence.

5 - Computerising mathematics assessment at GCSE: the challenge

On the papers analysed, just 16 out of 101 sub-tasks explicitly required method, working or reasoning as well as, or instead of, a simple result. Only one such question part insisted on seeing the method of a calculation which lead to a numeric answer – the other 15 asked students to explain or justify their reasoning²⁷.

Looking at the marks on a mark-by-mark basis, out of the 200 marks which were available:

- 102 of these could be awarded as partial credit even if the final answer to the sub-task was wrong or missing
- of these part marks, 77 were not “independent” as they would be awarded by default if the final answer to the sub-task was correct – so the main effect of method marks is as partial credit for students who make mistakes
- the remaining 25 part marks include the tasks which specifically asked for reasoning, and other “independent” marks that could be awarded for correct aspects of an incorrect response

It can be seen that partial and method related marks comprise a large fraction of the total available marks and, hence, eliminating some or all of these (to allow easy computer-based delivery and marking, for example) would have the potential to significantly affect score distributions. This raises a question: does the availability of these method marks actually add to the validity or fairness of the test, or does it add “noise” which could undermine the psychometric validity of the test?

The other question is how consistently such mark schemes can be implemented: the working may be illegible or ambiguous; the correct answer might appear in the working but not on the answer line; the working might show both correct and incorrect calculations or even reveal that the pupil obtained the “correct” answer using an incorrect method²⁸.

It was also noted that the GCSE mark schemes for those papers which allowed the use of calculators were still written with the assumption that most partial scores would result from arithmetic errors in the final calculation. Hence, on a typical question with two marks, the correct answer would automatically confer full marks, but to get a partial mark for an incorrect answer, the complete, correct expression for the correct answer would have to be seen in the students working. The main beneficiaries of such partial marks would, therefore, be students who eschewed the use of a calculator.

It would be informative to closely study the results of these tasks to see how often these partial marks were actually awarded (compared to those cases where the answer was fully

²⁷ In the new GCSEs piloted in 2010, an additional requirement has been added (to all GCSE subjects) that 3% of assessments should depend on “Quality of Written Communication.” Consequently, designated questions have one or two marks dedicated to clear communication, use of appropriate technical language, correct spelling etc.

²⁸ The *Triangle* task in Section 6.8 produced some examples of these issues.

correct or completely wrong) and whether there is an identifiable cohort of candidates who benefit from them. If these marks are not providing useful information then a case might be made that the complication they cause when adapting a test for computer delivery outweighs their usefulness.

However, the scoring of working and explanation is seen as vital to many of the progressive assessments discussed in Chapter 2. Such assessments also emphasise fewer, longer tasks featuring *extended chains of reasoning* in which the pupil must perform several steps, without prompting and support, to arrive at the answer. In such tasks, pupils may well perform several credit-worthy steps towards the solution before making a mistake. In some cases, only the higher performing candidates would be expected to reach a fully correct solution. Dropping the facility for method marks on the grounds that GCSE, with its emphasis on short 1 or 2 step sub-tasks, did not make good use of it could preclude the later introduction of longer tasks.

Heterogeneous testing

One noticeable feature of GCSE is that all of the questions are a similar form: relatively short, partially-constructed responses written directly onto the question paper (which typically provides a fixed space for writing an answer accompanied by space for showing working and reasoning). There are two papers of equal length – one allowing calculators, the other not, so there is little flexibility in the balance of calculator versus non-calculator questions. It is easy to see the practical and logistical reasons behind this: enforcing a calculator ban for just part of a test session would be chaotic, as would be mixing short answers on the test paper with fully constructed answers on plain paper. Multiple choice is most efficient when used with mechanical marking systems relying on “bubble” forms, encouraging multiple-choice only tests rather than a balance of question types.

Circa 1980, 'O'-level examinations included multiple choice, short answer and fully constructed answer questions, but this was at the expense of having a separate paper for each format and the risk of loose sheets of graph paper and extra booklets becoming separated.

Computer-based testing can, potentially, offer a more flexible solution. On-screen calculators can be enabled for part of a test and then automatically turned off. Every pupil gets the same calculator features, and invigilators no longer need to be experts at identifying proscribed devices with memory or internet facilities. Multiple choice questions can be freely mixed in with other types and still marked automatically, while other questions could be directed to human markers.

5.4: Adapting a GCSE-style paper task for computer

The design challenge

The World Class Tests project (Chapter 3) focussed on developing new task types specifically for computer, with only weak constraints on the curriculum to be tested. Where tasks were adapted from paper, the designers were free to make fundamental changes. The burden of adequately sampling the domain was shared with the paper-only tests and, even on the computer tests, paper answer books were available for any responses which could not easily be captured on computer. These luxuries are not enjoyed by the developers of high profile, high stakes tests such as GCSE: what might an online test developed to a more pragmatic brief look like?

If it were necessary to replace an existing, high-stakes test such as GCSE with an entirely computerised version (as the 2004 QCA announcement appeared to suggest) then an attractive solution might be to engage an IT company to adapt the oeuvre of proven, paper-based tasks and task-types to computer. Here, we discuss some of the design issues that such a process would raise, and how the conversion might change what the tasks assess (as was seen with some of the *Progress in Maths* tasks in Chapter 4).

The underlying issues were raised in section 2.5 - here we look at some more specific examples, with a particular focus on the question and response types observed in GCSE papers studied.

We conclude with a “worked example” which looks at various ways of presenting a simple paper task on computer.

Presentation issues

Quantity of information

In both the *Progress in Maths* and *World Class Tests*, the typical screen contained considerably less text and graphics than would typically be presented on a single sheet of paper. The Nelson paper tests regularly fitted two questions per A4 page, whereas the equivalent computer tasks used at least one screen per question – with a similar quantity of white space and illustrations. Longer questions were either split over two screens or adapted to use multiple column layouts.

Often, there is a shared table or diagram that needs to be reproduced on each page, further reducing the space available for each part of the question.

5 - Computerising mathematics assessment at GCSE: the challenge

It was noted during the *World Class Tests* development process that initial designs on one sheet of paper often required two or more screens to implement. Putting any more information on a single screen generally led to the design being rejected by reviewers as “too cluttered”.

There seem to be several influences behind this:

- A general tendency to find computer screens less restful or readable than paper.
- The resolution of computer displays still lags behind paper. One consequence is the increased use of larger, lower-density typefaces
- Computer screens use a “landscape” format, which invites the use of parallel columns of information. This may become more pronounced in the future with the increasing popularity of “wide screen” (16:9 or 16:10 aspect ratio) displays. Tasks designed on paper might not exploit this
- Assessment tasks have to be designed for the smallest display size and lowest resolution that might reasonably be found in schools²⁹ - this usually lags somewhat behind the “state of the art”
- Computer based tests invariably require on-screen controls and some extra instructions not present on paper, which all consume space on the screen
- At design review/approval meetings, tasks may be presented to a large group on a data projector. This may result in screen designs being evaluated as if they were *PowerPoint* slides, generally expected to contain a few bullet points, rather than being designed for individual use on a computer screen. The screen designs resulting from this process were quite sparse, even compared to other office or games software – but, equally, traditional examination papers are sparse compared to typical textbooks. More research may be needed as to the optimal amount of information which can be presented on screen.

Splitting a task across several screens could potentially affect its performance – especially where subsequent parts build on previous answers (rare, but not unknown, at GCSE) and pupils might need to refer back to a previous screen. Conversely, where a paper-based task consists of two independent questions, pupils might habitually skip the second part if they could not complete the first. Presenting such a task as two, separate, screens might encourage pupils to attempt both parts. In the nferNelson study (Chapter 4), there was a suggestion that scores increased (compared with paper) on the second screen of two-screen tasks.

²⁹ *World Class Tests*, circa 2000, aimed for 800x600 pixels on a 13” diagonal screen – this would be somewhat conservative, but still defensible, today.

Graphics, colour and multimedia

The design and presentation of GCSE tasks is highly conservative – black and white with simple, diagrammatic line drawings – which can be reproduced on the computer without technical difficulty.

A greater challenge would arise if, and when, it was decided to introduce colour, animation and other interactive elements. Experiences with some of the new question styles introduced in the nferNelson tests (Chapter 4) suggest that this can have repercussions, so it is essential that any such enhancements are driven by assessment goals and carefully considered by task designers rather than added by programmers out of a desire to use the technology.

Introducing colour also raises the issue of accessibility by the significant minority of students with colour perception issues. If it is decided (as with *World Class Tests*) that all questions must be accessible by colour-blind students as standard this severely restricts the use of colour. The nferNelson tests included a few items for which colour-blind students would require assistance (Section 4.5).

Some use of colour – or at least shade – is needed to compensate for the different nature of paper and computer displays. Fine, black “hairlines” on graphs, for example, cannot always be made sufficiently thin on screen – the solution is to use grey or a pale colour.

The consistency of the display also needs to be ensured, in terms of clarity, aspect ratio (circles must appear circular) and colour reproduction (which varied enough between machines to raise some issues with the nferNelson tests). For high-stakes testing, standards for display equipment would need to be established and monitored.

Using any form of sound means that headphones have to be supplied and maintained, and provision has to be made for pupils with impaired hearing.

Response gathering and marking

There are two particularly important constraints on the way the computer captures the candidate's response to each question:

- To enable automatic marking, the responses must be captured in a well-defined, unambiguous format that can be easily and reliably interpreted. Developing sophisticated “artificial intelligence” systems to interpret (for example) freehand diagrams would be possible, but expensive
- The candidate must be provided with a “natural medium” in which to respond to the problem (see section 2.5). Operating the computer should not detract from the mathematics. Where candidates need to master significant ICT skills before the test,

5 - Computerising mathematics assessment at GCSE: the challenge

these should ideally be transferrable skills with long-term value, not specific to the testing system in use.

Numbers

It is unsurprising that a large swathe of mathematics tasks can be answered with a simple number: returning to the pair of GCSE papers analysed in depth, 55% of the sub-tasks yielded answers in the form of integers, decimal fractions or probabilities and 32% of the individual marks were obtained directly from these responses. An equal number of method marks were awarded automatically where these answers were correct.

Some care is needed when marking numerical responses, especially decimal fractions. Direct equality tests on “floating point” numbers can be unreliable in some software environments as small rounding errors are common³⁰, so rather than asking “is the answer equal to 0.1” it is safest to ask “is the answer between 0.9999 and 0.1001”). It is also possible that, in some tasks, leading or trailing zeros which make no numerical difference might be important (for example, “£0.5” might not be acceptable in place of £0.50 in a realistic money question, or “3.100” might wrongly imply that an answer was accurate to three decimal places). When such issues arise on paper, they can be raised and addressed during marker training, marking or moderation. For a computer-based system they need to be foreseen and the criteria for acceptance specified in advance.

Mathematical Expressions

Apart from questions that involved drawing graphs or diagrams, about 20% of the marks still required the student to write down answers that could not be represented as “plain text”.

These include:

Fractions & division expressions : “One half” can reasonably be written as “1/2” but “one and a half” gets a bit messy as “1 1/2” and could easily be misread/mistyped as “11/2”. In expressions such as:

$$\frac{3+5}{2}$$

...the notation includes implied information about the order of operations (3 + 5 /2 would conventionally be evaluated as 5.5, not 4) and would need to be entered as (3+5)/2 adding “understanding brackets and order of operations” to the assessment aims of the question.

³⁰ A consequence of the binary representation of real numbers is that some results require rounding to the nearest “binary place” - analogous to writing $\frac{2}{3}$ as 0.6667 in decimal. The resulting errors can accumulate in calculations and so tests for **exact** equality may fail unexpectedly.

5 - Computerising mathematics assessment at GCSE: the challenge

However, perhaps surprisingly, only two sub-tasks on the papers analysed required a fraction as the final answer, with two additional part marks depending on spotting fractions in the working.

Powers and Standard Form – this includes expressions such as “x²” (which would be “x^2” in most spreadsheets or “x**2” in FORTRAN) and standard form/scientific notation such as “ 1.5×10^3 ” – typically written as “1.5E+3” on a computer. A purist might argue that this is not “standard form” – it is certainly less descriptive than the traditional notation.

Multiplication: There is no “times” sign on a computer keyboard. Typists would use “x” (ambiguous in an algebraic expression), spreadsheets and computer languages usually use “*”. In written algebra multiplication is often implied (“5n”) which is something that auto-marking would have to consider.

The simple answer is to rely on the conventions established by spreadsheet software – but a complete abandonment of traditional mathematical symbolism in a major exam would surely be controversial. In some cases it would be possible to provide a “template” for the answer – e.g. separate boxes for the numerator and denominator – but in other cases (such as “express 123460000 in standard form”) that would undermine some of the purpose of the question.

Measuring and Drawing

7% of the marks on the sample paper required the drawing of graphs and diagrams. Usually at least one question also requires the use of actual “instruments” (ruler, protractor, compasses) to measure quantities and construct diagrams. Should a computer implementation of these try to “simulate” the physical instruments (e.g. a virtual “protractor” that can be dragged around the screen) or should it provide something more like a CAD or interactive geometry package? The latter would require a substantial change to the syllabus – but would represent a more realistic and transferable use of ICT and of modern functional mathematics.

Any graph or drawing based answer will require candidates to master some sort of – potentially unfamiliar – user interface allowing them to draw, adjust and delete lines and points. Marking such responses also needs more sophisticated analysis than simply checking against a correct answer.

Rough work and partial credit

Virtually **every** question included several lines for “rough work”, and 27% of the total marks could potentially be awarded for evidence of correct reasoning.

5 - Computerising mathematics assessment at GCSE: the challenge

For example, in a question about calculating average speed from a graph, the final answer is a simple, easily auto-marked number worth 3 points, but if the student gets this wrong, partial credit is awarded as follows:

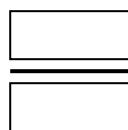
“200 miles” and “2½ hours” (or equivalent) seen in work: award **1 mark**

$\frac{\text{(their 200)}}{\text{(their 2½)}}$ seen in work: award **1 mark**

So, the student gets at least 1 mark for dividing distance by time, even if they use the wrong values – but obviously shouldn’t be rewarded for invalid work such as dividing time by distance.

The first issue for eAssessment design is capturing this working. A question might have an easily captured and marked final answer but the working might still involve awkward-to-type expressions. The problems of entering mathematical expressions are discussed elsewhere, but the need to present candidates with a “natural medium” for working is particularly important in this context.

Any sort of “structuring” to the working space – such as:



... gives the student extra information on how to answer the question.

A plain text box is not a comfortable medium to work out calculations for which the traditional layout often serves as a visual aid to calculation.

So, if the expression $\frac{200}{2\frac{1}{2}}$ were typed as “200 / (2 1/2)” the step of multiplying through by 2 to remove the ½ becomes much less obvious.

One possible approach is to abandon the idea of capturing working as it happens, and make clear to the student that presenting their method in computer-readable form is an additional, required step which they should perform after solving the problem.

Automatic marking of correct reasoning presents further problems. There is no “right answer” in this context: markers are trying to give the benefit of doubt to the candidate. Judging whether the student knew they were dividing distance by time may be possible for a marker with knowledge of the question and experience of common mistakes made by students, but expressing that as a computer algorithm is more of a challenge.

Translating a sample task

Let us suppose that a substantial part of the current examination is to be turned into a computer-marked, on-screen test, and that a substantial re-design of the syllabus and question style is not part of the designer's brief.

Some questions, such as those with graphical responses – will obviously pose technical challenges – but what about the apparently simple questions?

Here we consider a simple, short question of a type that might appear on a maths examination paper:

1. Put the values 0.236, 0.4, 0.62, $\frac{1}{5}$, $\frac{3}{8}$ in order, smallest first

.....
.....

Answer: (2 marks)

What follows is a commentary on some of the design decisions that arise – especially surrounding the style of response - and the implications they may have for the equivalence of the task to the paper version. Such implications might easily be overlooked by a designer or programmer with limited experience in mathematics education.

Multiple choice

The easiest type of response to capture and mark on computer is the traditional multiple choice question. So, the question could be re-written as:

2. Which of these shows the numbers in order, smallest first?

(a) 0.4, 0.62, 0.236, $\frac{1}{5}$, $\frac{3}{8}$

(b) 0.236, 0.4, 0.62, $\frac{1}{5}$, $\frac{3}{8}$

(c) $\frac{3}{8}$, $\frac{1}{5}$, 0.236, 0.4, 0.62

(d) $\frac{1}{5}$, 0.236, $\frac{3}{8}$, 0.4, 0.62

(e) $\frac{1}{5}$, 0.4, 0.62, 0.236, $\frac{3}{8}$

This has increased the amount of material which the student has to read, and means that the student will spend time analysing the alternative answers rather than performing the sequencing activity. Choosing the “distractors” also needs careful thought and research to ensure that the question still probes the original assessment goals as opposed to good “exam technique”. While the original required each value to be correctly placed, in the example above, it is sufficient to realise that “ $\frac{1}{5}$ ” is the smallest and that 0.62 is bigger than $\frac{3}{8}$. Well-

5 - Computerising mathematics assessment at GCSE: the challenge

crafted multiple choice questions can be effective assessment tools, but they are not necessarily “drop in” replacements for constructed answers.

Alternatively, one might change the question:

3. Which of these numbers is the **smallest**?

(a) 0.236 (b) 0.4 (c) $\frac{1}{5}$

(d) 0.62 (e) $\frac{3}{8}$

This is straightforward, easy to read and tests a very specific curriculum point without entangling it with other topics. However, this is clearly not the question we started with – it is only asking for the smallest number – students might just chose $\frac{1}{5}$ “because fractions with 1 on top are small” and still get the mark. With a large battery of multiple choice questions, the effect of guessing on the final score may be within acceptable limits, but it would still be unsafe to make inferences about any individual response. This becomes important if the test is to have formative or diagnostic value. In contrast, the answer to the original question could expose a number of common misconceptions such as “0.62 is smaller than 0.4 because its in hundredths” or “0.246 is bigger than 0.62 because 246 is bigger than 62”. Several short multiple choice “items” each focussed on one of these misconceptions would be needed to reveal the same information.

So, how could the question be implemented without converting it to a multiple choice “item”?

Typed responses

The original question could be used, along with a set of type-in boxes for the answers:

1. Put the values 0.236, 0.4, 0.62, $\frac{1}{5}$ $\frac{3}{8}$ in order, smallest first

Space for working

Answer:

Smallest					Largest

That seems acceptable – but how will the student enter $\frac{3}{8}$ in the answer box? If “3 / 8” is acceptable, how would that extend to other questions involving mixed numbers? How and when will the student learn these conventions?

The “space for working” has been included as it was present in the original question: is this still useful – either as a natural way of working for the student or as a mechanism for awarding partial credit? What is the pupil expected to enter there?

What other minor variations of answers need to be specified to the computer? For example, what if the student types “.0236” in place of “0.236”? Had this answer been the result of a calculation, then the mistake would be unacceptable, but in this context it is clearly a typing error which might be overlooked. A human marker only needs a general instruction such as “accept obvious and unambiguous transcription errors” - for a computer-marked test the exact rules for such allowances would need to be codified.

From a perspective of user-interface design, it is often helpful to constrain or automatically correct inputs at entry time: for example, it could be made impossible to type in an invalid number (e.g. with two decimal points); leading zeros could be automatically filled in (i.e. “.4” would automatically become “0.4”) or trailing zeros stripped (“0.40” becomes “0.4”). In this task, these features would probably be helpful but in another question (e.g. “round 0.395 to one decimal place”) “0.40” rather than “0.4” would be a poor answer which the examiner might wish to penalise. Hence, such detailed design decisions need to be made, on educational grounds, by the question designer and not by a software designer whose sole consideration was ease-of-use.

Drag and Drop

Most modern eAssessment authoring systems could easily cope with the following:

1. Put the following values in order, smallest first.

(Drag the numbers into the boxes below.)

0.236	0.62	0.4	$\frac{3}{8}$	$\frac{1}{5}$
---------	--------	-------	---------------	---------------

Smallest

Largest

--	--	--	--	--

This eliminates any worry about mistyping or using the wrong convention to enter a value. There are a few details to specify: can you drop the same number in more than one box (not here, but it might be appropriate in other tasks)? How do you undo mistakes? Do the instructions for doing this need to be on the screen, or can it be assumed that pupils know how to operate such “drag and drop” interfaces?

This seems to be a reasonably faithful “clone” of the original question – but it is still a matter for debate – or experiment – as to whether the question is completely equivalent in difficulty to the original. For instance, it is now easier for the student to use “trial-and-review” – juggling the numbers until they are satisfied with the answer. This may actually result in a better task, but changing the presentation in such a substantial way is liable to invalidate any existing calibration data on the original question (see Bodin, 1993).

Partial credit

If there is any partial credit to be awarded then what are the precise rules? The correct answer is:

$\frac{1}{5}$	0.236	$\frac{3}{8}$	0.4	0.62
---------------	---------	---------------	-------	--------

If the mark scheme rule is “2 marks for a correct answer; 1 mark for a single misplaced number” then how should the following be interpreted:

$\frac{1}{5}$	0.236	0.4	0.62	$\frac{3}{8}$
---------------	---------	-------	--------	---------------

A human marker might reasonably say that there is just one mistake: $3/8$ is in the wrong place. However, a simple marking algorithm would register three mistakes (the last three boxes all contain the wrong numbers). The testing system would have to be capable of supporting the more sophisticated task of counting misplaced items in a sequence, and the task designer would have to specify that the question required this behaviour – which might not be appropriate in other tasks.

The best solution?

In this case, “drag and drop” seems the closest match to the original paper task (although a multiple choice version with more carefully crafted distractors than the deliberately poor example shown here might also work) – but a slightly different question (e.g. “Convert these fractions to decimals”) might suggest a different style. The key, however, is that the best decision will be one that is informed by knowledge of students' common mistakes and misconceptions – and that requires a designer with a combined ICT and mathematics education background or a close collaboration between experts in each field.

Whether the final question is truly equivalent to the original paper one is best decided by controlled trials – although ultimately this may only reveal whether the new question is “equally difficult”, not whether it tests the same mathematics.

5.5: Conclusions

New demands on assessment designers

Research questions B and D asked “what are the effects of transforming an existing paper-based test to computer?” and “what does this imply for the technical and pedagogical processes of computer-based assessment design?” This chapter has explored those questions by setting the imaginary task of producing a computer version of a well-established high-stakes test, and taking the first steps in the design process.

The detailed points discussed in the previous section illustrate the type of issues that might arise when designing or adapting assessment tasks for computer, and which might not occur to experienced designers of paper-based assessment. In most cases, although the issues arise from technological considerations, the decisions depend on understanding the assessment aims of the task and the types of mistake and misconceptions it might expose.

In summary,

5 - Computerising mathematics assessment at GCSE: the challenge

- Adapting even a simple paper question for computer delivery and marking can require a number of detailed design decisions, encompassing both technical and pedagogical issues, which could affect the performance of the task
- Task designers will need some knowledge of the technical capabilities of the delivery and marking systems, and will need to specify tasks to a greater level of detail than is normal when drafting paper assessments
- A programmer may have to modify the task to fit the system, but may not be qualified to make the correct decisions from an assessment perspective
- Unless designers are skilled in both assessment design and technology, they must work in close co-operation with those responsible for implementing their tasks on computer. The “book publishing model”, in which the task designer writes a manuscript and hands it over to a publisher to be typeset, with one or two opportunities to review proofs, will not work here. There needs to be a clear, equitable communication channel between task designer and programmer, supported by people with cross-disciplinary experience.
- Software for delivering tests needs to be designed to support the aims and needs of mathematics assessment. Existing systems designed for general assessment, with an emphasis on simple, short answer or multiple choice questions, or even a sophisticated text analysis system for marking long written answers, may not have the flexibility to deal with the particular demands of mathematics.

A critique of GCSE Mathematics

As part of research question C: “How might eAssessment be used to improve the range and balance of the assessed curriculum?” it is also reasonable to look at some of the shortcomings of the “state of the art” of GCSE, and how it might be improved upon. The design analysis of tasks conducted here provides some insight on this. Indeed, the process of considering how a task might be translated to computer leads naturally to analysing the design of the original task to identify the key assessment objectives.

Some features of GCSE, particularly the “constructed response” style of question and the widespread awarding of marks for partially correct working, would be challenging to replicate faithfully on computer, so it is reasonable to raise questions about the value of these features.

There is a need for caution here: it is tempting to conclude that GCSE fails to make consistently good use of constructed responses and could be quite satisfactorily, and more cheaply, replaced with a mixture of short answers and multiple choice. This could preclude

5 - Computerising mathematics assessment at GCSE: the challenge

future improvements in assessment which would rely on such features to (for example) set more open questions requiring extended chains of reasoning, for which capture and scoring of working would be vital.

One overarching question, though, is whether many of the traditional task types are relevant for an ICT-driven world. Examples of conventions which seem incongruous when presented by a powerful calculating machine include

- partial marks which only come into play when the candidate makes arithmetical errors, when a calculator is available
- estimation tasks which focus solely on approximate calculation (often using rather contrived expressions) rather than other techniques, such as predicting orders of magnitude, which are useful for checking the plausibility of calculated results
- manually calculating summary statistics for tiny, often special-case, data sets rather than interpreting the meaning of these measures in the context of large, realistic data sets. Learning to blindly apply techniques which can not be safely applied without a deeper understanding of the use and abuse of probability and statistics
- drawing graphs and geometric constructions by hand
- “heterogeneous” tests (e.g. all short constructed answer, separate calculator and non-calculator papers) when a computer could offer a mix of response types and allow/disallow calculator use on a task-by-task basis

In the following chapter, some of these issues and questions are put to the test in the design, development and evaluation of a prototype “GCSE-like” computer based test.

6: Design and evaluation of a prototype eAssessment system

/ You are not expected to understand this */*

*Comment in the source code for the
Unix Operating System (6th Edition)*

6.1: Introduction

The *World Class Tests* project (Chapter 3) showed how ambitious new types of assessment task could be presented on computer. However, this development was not constrained by the need to deliver an existing curriculum or maintain comparability with established tests. It relied on extensive programming to implement each new task and a resource-intensive combination of data capture, written answer booklets and manual marking to deal with the variety of student responses. The *Progress in Maths* tests (Chapter 4) showed how an existing test could be presented on computer, but in this case the existing test was largely dependent on short, simple answers, so the issue of how to translate “constructed response” questions was not tackled.

Consequently, it was decided to develop a prototype system with which to explore whether GCSE-style questions which use “constructed responses” or give credit for working could be successfully computerised. This system could then be used in small-scale school trials to evaluate these tasks and compare their performance with the paper originals. Additionally, for comparison, the system could deliver “simplified” versions of the tasks, reduced to simple numerical or multiple choice answers. A key question (see 5.4) was how much real

6 - Design and evaluation of a prototype eAssessment system

information could be reliably gleaned from the extra response data, and whether this justified the technical effort.

In the light of previous experience, particularly of ad-hoc systems used during the trials of the *World Class Tests* materials, the main requirements of the system were specified as:

1. The ability to rapidly assemble questions from standard “components” without significant programming
2. The flexibility to add new “components” enabling novel question types to be evaluated
3. The ability to capture responses in an easily processed format (XML)
4. The facility for students to move back and forth through the test questions to review and possibly correct their answer (this seems obvious, but is not straightforward especially with questions with complex interactive elements)
5. The ability to reconstruct the student's responses, as seen on-screen by the student, for presentation to human markers
6. Easy installation for trial schools, including the possibility of running the entire test in a standard web browser with common media plug-ins
7. Data to be saved over the internet to a central server as the test progresses. This decision was based on experiences with World Class Tests which showed that writing data to local storage and subsequently collecting it caused major problems. By the time of this study, it was reasonable to require that participating schools had a good broadband internet connection: this would have been unrealistic at the time that World Class Tests were initially conceived
8. Some degree of crash resilience, such as the ability to resume an interrupted test, possibly on a different machine
9. Where possible, use open-source products, or other freely available systems. The end users should not be required to pay for any software, and if possible the need for expensive proprietary software (such as database managers) on the central server should be avoided
10. Potential accessibility for students with special needs: although this did not need to be fully implemented for the trials, there should be “proof-of-concept” that it could be achieved
11. Cross platform compatibility (PC/Mac/Linux) was desirable but not essential. As with the previous point, this need only be at “proof of concept” level, to show that

such a system could be implemented without mandating a particular proprietary operating system

The overall system is described in Appendix A. In this chapter we will summarise some of the less conventional features of the design, including two input tools which attempted to expand the capabilities of the system beyond simple short answers, and the hybrid, online marking system.

6.2: Some general-purpose tools

Although the project could not hope to provide a complete solution to the design challenges above, it focussed on a small number of techniques each of which addressed issues raised by a substantial “genre” of GCSE task types. The two main issues addressed were the capture of “method” in calculation tasks (with potential application to the majority of GCSE questions) and the plotting of points or straight lines on graphs.

For this study, a priority was to make the tools simple enough for pupils to use without prior experience. If such tools were to be adopted in a live examination, it would be reasonable to expect pupils to be trained on the tools in advance, so slightly more sophisticated features and options could be considered.

The “Printing Calculator”

Chapter 1 raised the question “is the computer a natural medium for doing mathematics?” and pointed out the lack of generic mathematical tools with which students' fluency could be assumed, in contrast with the ubiquity of the word processor as a tool for writing. This poses a problem in a mathematics task in which credit is given for correct or partially correct working: the instruction “show all your working” assumes that the response medium is suitable for working.

One mathematical tool with which most students are familiar is the pocket calculator. An on-screen calculator which recorded the sequence of operations might be a viable way of capturing method in questions which could be reasonably “worked out” on a calculator.

It would be possible to store every keystroke made on such a calculator without the user being aware – but this could include errors and “blind alleys” which the student had tried and rejected. Any automatic marking system would face the non-trivial challenge of reliably inferring which keystrokes truly represented the student's thinking.

Instead, it was decided that the student should have the final say of how their working was presented. The proposed solution was to simulate a calculator which produced a “till roll” printout. As an integral part of the task (emphasised in the initial instructions) the student is

6 - Design and evaluation of a prototype eAssessment system

asked to “show their working” by dragging a strip of printout from the calculator and “pinning” it to their answer, in addition to typing in the final result.

Triangle

ABC is a triangle.
 $CA = CB$.
 D is a point on AB

Not drawn accurately

(a) Work out the value of x .

Answer: $x = 55$ degrees (2 marks)

(b) Work out the value of y .

Answer: $y = 20$ degrees (2 marks)

Drop the printout from the calculator here to show your working

Task 2 of 7 | Back | Page 2 of 14 | Next | Quit

Figure 6.1: The “printing calculator” concept (the large arrow represents the student dragging the “printout”)

Figure 6.1 illustrates what the student sees and does. Figure 6.2 shows how a correct response to this question would be presented to the marker, along with the mark scheme. Alternatively, as the calculation steps are stored in a well-defined format, algorithms can potentially be developed to automatically mark the working. The example shown (which mimics the mark scheme for a common GCSE task) illustrates the issue raised in the previous chapter about the redundancy of partial credit for faulty arithmetic: if “180-55-55-50” is “seen” it is most likely that the correct answer of “20” will also appear. Given that everybody would answer this using a calculator, partial credit could instead be awarded for sight of an intermediate result such as 110 or 70.

6 - Design and evaluation of a prototype eAssessment system

2.1 180-55-55-50 seen	M1 <div style="border: 1px solid black; padding: 5px; margin: 5px;"> $55+55=$ 110 $180-110=$ 70 $70-50=$ 20 </div>	1
2.2 20	A1 Answer= 20	1

Figure 6.2: Student response using the printing calculator, with working, presented to marker

The calculator functionality chosen was loosely based on that of a popular “scientific” model, marketed at GCSE and A-Level students, in which the traditional 10-digit numerical display was supplemented by an alphanumeric display showing the current calculation. As is common with such products, the on-screen calculator applied the correct ‘order of operations’ to calculations such as “ $3 + 5 \times 2$ ” and understood brackets. Fractions and standard form could be entered and manipulated using the same input conventions as a “real” calculator although, taking advantage of the computer's display, the presentation of these was improved slightly. As well as capturing working, the calculator could be used as a way of answering questions with fractions or standard form.



Figure 6.3: Advanced uses of the calculator

6 - Design and evaluation of a prototype eAssessment system

The range of features implemented represents a compromise between a basic “four function” model and the large number of functions found on “scientific” models, many of which are unnecessary at GCSE level. The functions included were more than adequate for the task set used in the evaluation. To implement a full, intermediate-tier GCSE curriculum, basic trigonometric functions could easily be added, while a more comprehensive review of GCSE specifications and task types would be necessary to determine the exact set of functions needed for the higher tier. This study concentrated on intermediate-tier papers.

Some common features of calculators were deliberately omitted. Percentage keys on calculators conventionally work in idiosyncratic, non-algebraic ways³¹; it was also felt that the meaning of percentages was something with which pupils should be familiar. Memory functions could have made the “printouts” harder to interpret, especially if the memory contents were carried over from a previous printout.

Potentially, several “levels” of calculator functionality could be implemented and the task designer could specify which features were to suit the requirements of each task, although this risks giving away clues as to how to answer the question. It might be sensible to maintain the same calculator design throughout a test.

Prohibiting the use of calculators

The GCSE examinations under consideration comprised two papers, only one of which permitted the use of calculators. Both papers require students to show their working, which is particularly significant on the non-calculator paper where arithmetic errors in otherwise-correct working seem more likely (see Section 5.3). Clearly, the “printing calculator” technique for capturing working could only be used where calculators were allowed.

The prototype system allowed the calculator to be enabled and disabled for each individual question. In a real examination, provision would have to be made for preventing students cheating by using of the calculator provided in one question to work out the answer to another. One scenario is that the calculator would not be accessible until after the “non-calculator” portion of the test had been completed and the responses irrevocably saved. This would still be considerably more flexible than the current system in which calculators can only practically be permitted or prohibited for an entire test sitting, and could mean that the non-calculator portion of a computer-based test could be shorter and focussed on testing basic skills.

31 For example: even on a Casio calculator which otherwise uses algebraic notation, and will correctly evaluate expressions such as “ $100+2x3=$ ”, the sequence to add 5% to 200 is “ $200 \times 5 \% +$ ”.

The expression writer (proposed)

Another solution to the “no calculators” issue would be a version of the calculator that was purely an input device for illustrating working, and could not evaluate expressions.

This also has a potential use in responding to questions in which the answer may be a fraction, or best expressed in standard form (e.g. 1.3×10^8) since there are well-established ways of entering these on a calculator, with which many pupils should already be familiar.

Adding buttons with symbols such as “ x ”, “ y ”, “ a ” and “ n ” would allow the entry of responses to questions such as “write down an expression for...” However, it seems unlikely that such a system would feel natural enough to use as a working medium for a multi-step algebraic manipulation (such as factorising a quadratic).

The “expression writer” tool has not yet been developed further – partly because of time constraints but also because it was felt almost inevitable that candidates would have to be trained in advance to use such a tool.

The graphing and drawing tool

This was designed to enable a range of response types which required:

- Drawing line graphs
- Plotting points
- Plotting lines and points over fixed data (e.g. “line of best fit”)
- Labelling points, lines and regions
- Simple diagrams (containing straight lines)

The solution chosen was a fairly simple tool which allowed the task designer to specify axes and scales (or a plain, squared grid), onto which the candidate could draw points (by clicking) and lines (by clicking start and end points). The graph can have a pre-defined data set plotted, which the student can draw over. Draggable labels were implemented, so that the candidate can identify particular lines and points.

6 - Design and evaluation of a prototype eAssessment system

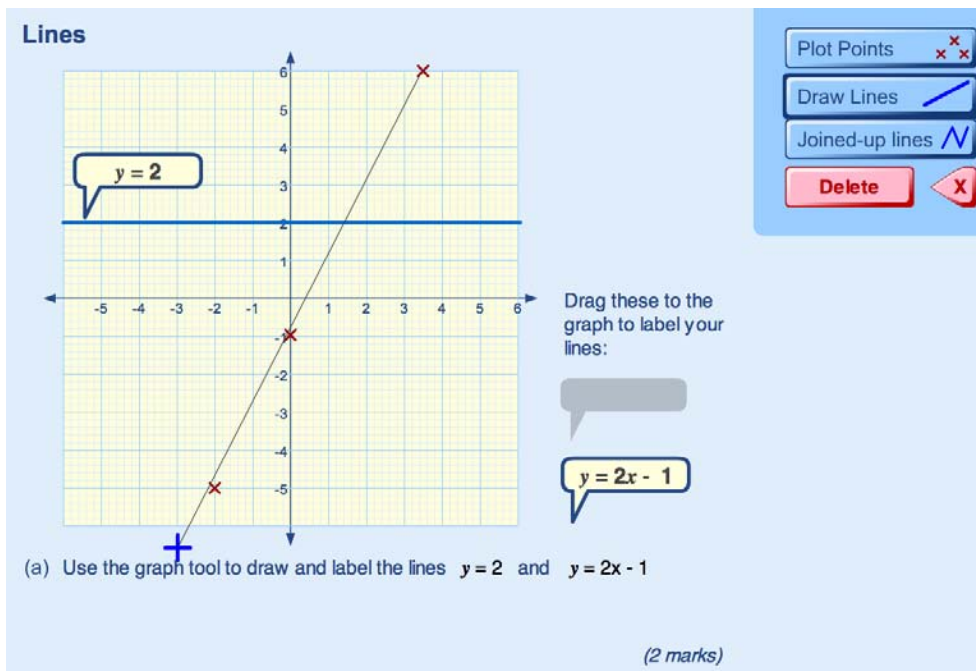


Figure 6.4: The line-drawing tool

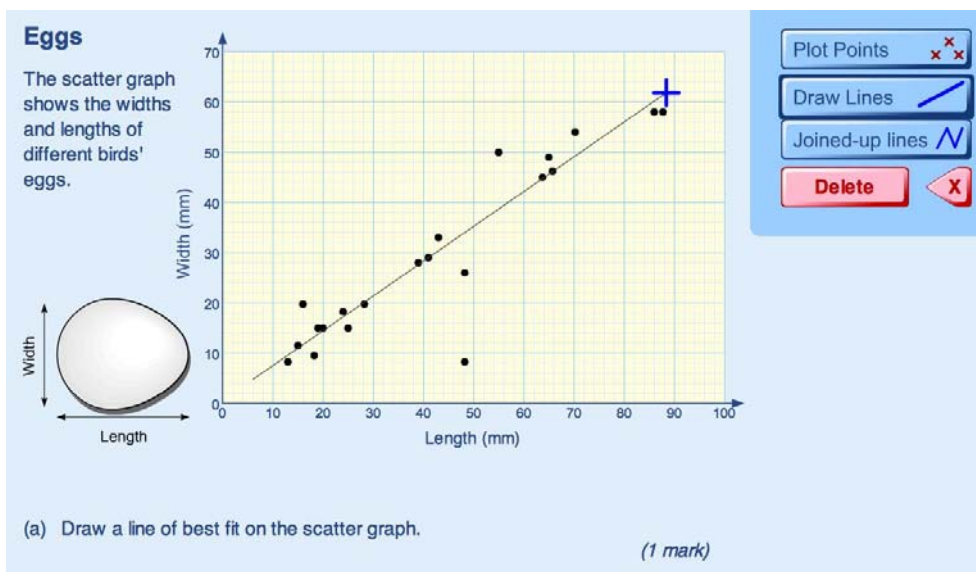


Figure 6.5: Drawing a line of best fit

Figure 6.4 shows one way in which the new medium might alter the nature of the task. The usual “paper” method would be to plot 2-3 points – preferably easy-to-calculate ones, then use a ruler to draw a line through them, extending to the edges of the graph. A strict mark-scheme might penalise a line which simply joined the points and didn’t extend to the edges.

6 - Design and evaluation of a prototype eAssessment system

Here, though, a line is defined by clicking on the start and end points – you can only see the line “move” when you have selected and fixed the first point. So, the optimum method of drawing a line from edge-to-edge is to work out the points where the line hits the edge of the graph – which involves solving the equation for given y , rather than substituting for x .

This is also evident in Figure 6.5, which shows the tool being used to draw a “line of best fit” over a set of pre-defined data points. Here a “physical” ruler is a useful tool, since it is possible to adjust the ruler to get a visually satisfactory fit before drawing the line. The on-screen tool could place greater demands on the student's ability to visualise the line before picking the starting position and, again, could encourage them not to extend the line beyond the data points.

One solution would be to simulate an on-screen ruler, which could be quite cumbersome (you would need a method of rotating it as well as dragging it around) and it is doubtful if this would really replicate the “real world” ease of use. Alternatively, once the line had been drawn, the end-points of the line could be draggable, as in a typical computer drawing package, but still only one endpoint could be moved at a time, and some students might not realise that such a facility was available. It was decided to persist with the very simple system shown here.

The source of this dilemma is that the “natural” way of performing these tasks on a computer would be to use a graphing or geometry package which could automatically plot a line or calculate a least squares fit more accurately than could be done by hand; so the true solution would be to change the curriculum to include authentic tasks which assessed the use of graphing and statistics tools to solve a problem rather than simply testing the ability to plot a graph or draw a line of best fit.

The student responses are stored as co-ordinate data, so it is possible to construct algorithms which will automatically “mark” the graphs. Figure 6.7 shows a completed graph: the column labelled “C” shows that the computer has been able to mark this question correctly.

6.3: Online marking tools

During the design of items for the *World Class Tests* project, an “in house” online marking system had been developed to enable trials of computer-based items to be scored, before the final delivery and marking system (developed by a third party) was available. Although the prototype was never “scaled up” to a state where it could be used for marking the live tests, it did have several features not found in the final *WCT* system, such as the ability to re-display the screen as left by the candidate rather than simply presenting the numerical responses.

Some of those ideas were incorporated into the design of the more sophisticated system developed for these trials.

As with *World Class Tests*, the study suggested a “hybrid” marking approach in which the computer automatically marked what it could and human markers completed the job, optionally reviewing and correcting the computer marks in the process. Alternatively, to provide a baseline for comparing human and computer marking, the entire test could be manually marked by the human marker.

The candidates' responses were reproduced alongside the text of the mark scheme, which used similar notation to traditional GCSE mark schemes. Figure 6.6 and Figure 6.7 show examples of responses using the printing calculator and graphing tools as they would be presented to human markers. In Figure 6.6 the computer has successfully allocated a part mark by spotting 2100/7 in the working³².

The same software was used to record the scores on each question part from the paper trials, so that they could be analysed alongside the computer results.

³² The question was: to make concrete you mix cement, sand and gravel in the ratio 1:2:4. How much sand do you need to make 2100kg of concrete?

6 - Design and evaluation of a prototype eAssessment system

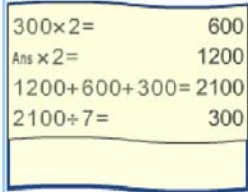
Marking:	1	School:		Start:	02:35	Date:	24 Nov 005
Student ID:	327			End:	02:58	Time:	12:23:30
2) Concrete				Attempt 1 of 1			
Markscheme		Response	C	M			
1.1	2100/7 (=300)	M1 	1	<input type="checkbox"/>			
1.2	600 (kg)	A1 Answer = 300kg of sand	0	<input type="checkbox"/>			

Figure 6.6: The marking system, showing computer-allocated partial marks using the printing calculator tool. Here, partial credit has been given for the sight of 2100/7 in the working.

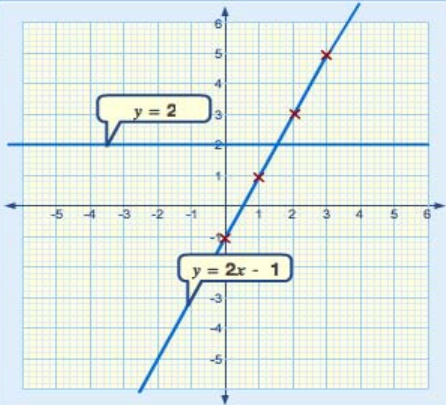
Marking:	1	School:		Start:	03:04	Date:	24 Nov 005
Student ID:	335			End:	03:26	Time:	12:21:20
4) Lines				Attempt 1 of 1			
Markscheme		Response	C	M			
1.1	Line $y=2$ through 3 correct integer points		1	<input type="checkbox"/>			
1.2	B3 Line $y=2x-1$ on graph through 3 correct coords B2 Any line with gradient 2 or 3 correct points plotted B1 Any line through (0,-1) or 2 correct points plotted		3	<input type="checkbox"/>			
2	Correct coordinates of intersection	Bft1 Lines cross at: $x=1.6$, $y=2$	1	<input type="checkbox"/>			

Figure 6.7.: Marking a graph question, showing automatic marking

6.4: Analysis tools

The system included an online tool for exploring the results of the computer and paper trials.

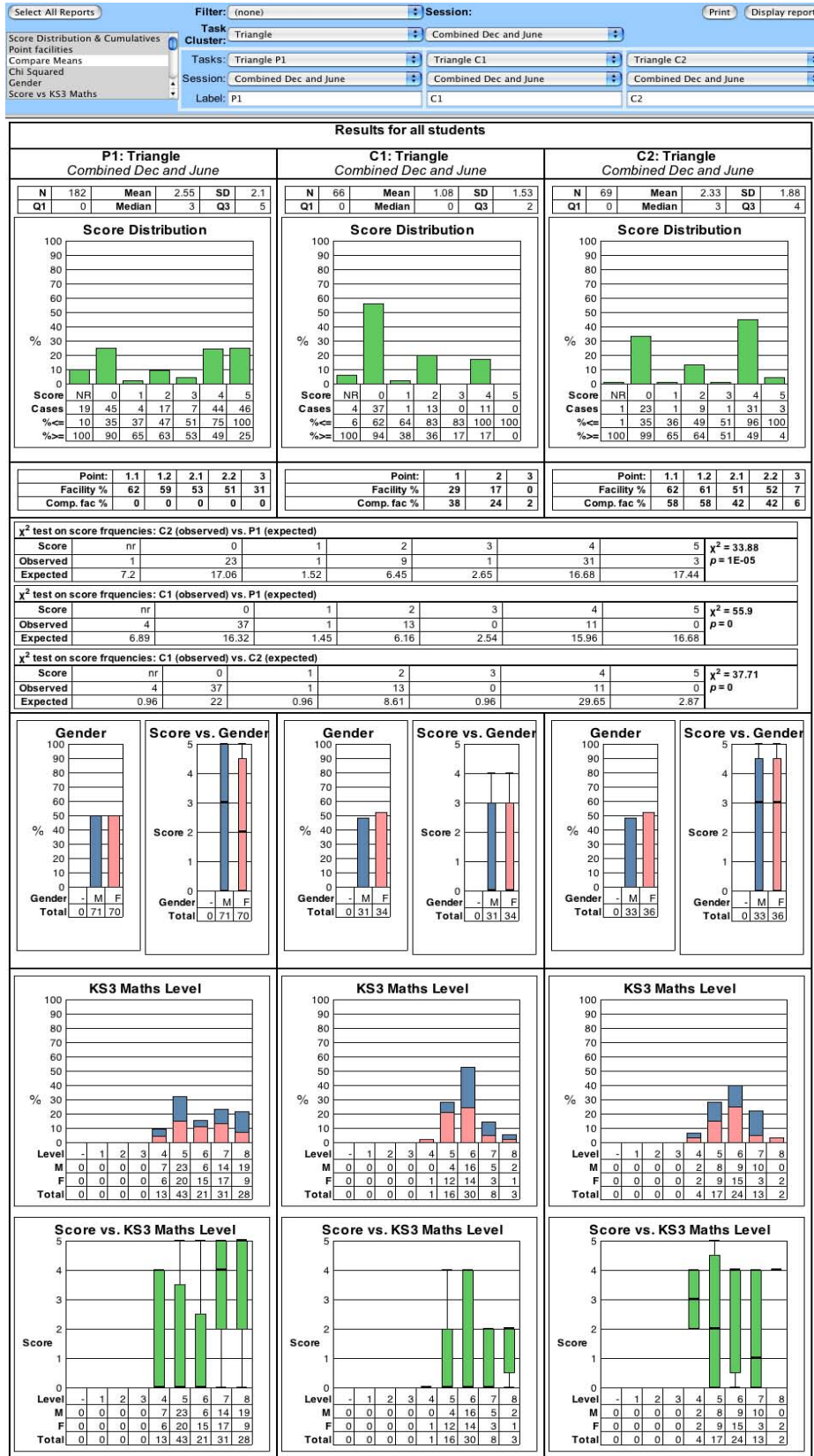
This could generate a range of statistics and distributions:

- score distributions & cumulative scores
- facility levels on each mark scheme point
- compare distributions between task variants using Pearson's chi-squared. This is to be treated with caution given the known limitations of the data and the small numbers (see page 158).
- sample composition distributions and score box-plots vs. gender, Key Stage 3 mathematics level, predicted GCSE scores and KS3 English levels
- Facilities for filtering the data by KS3 maths level and other criteria

A sample of this output is shown in Figure 6.8.

More data was gathered on the system for possible future analyses – such as timing information, point-for-point breakdowns of scores and the marks awarded by automatic marking.

6 - Design and evaluation of a prototype eAssessment system



6.5: The trial study of GCSE-style tasks

Aims

The primary research question for this part of the study was “what are the effects of transforming an existing paper-based test to computer” and “what tools and techniques might assist with this process?” The *Progress in Maths* study explored this in the case of comparatively short questions for young pupils: here we are looking at longer questions requiring more sophisticated mathematics. A secondary question was whether modest improvements in range and balance of tasks might be realised on computer (as distinct from the more radical departures of the World Class Tests project).

The task set

A selection of generic GCSE Intermediate Tier task types were identified, and “original” paper-based variants were written, closely matching the GCSE tasks in style and content. This process is not dissimilar to the way similar tasks types are re-used year after year (this became clear during the analysis of past papers in the previous chapter). In addition, a few less-typical tasks were adapted from the *Balanced Assessment in Mathematics* (Section 2.3) and *World Class Tests* (Chapter 3) materials.

Each of these was then used to design a “cluster” of (typically) 3 variants of each tasks, using different design approaches but with the same context and assessment objectives.

Variant P1: The conventional task answered on paper to act as control

Variant C1: Computer-presented task, re-worked to avoid the need for “rich” response capture (e.g. multiple choice, or no capture of “method”) so that it could be easily implemented on most e-assessment systems

Variant C2: Computer-presented task, re-worked to use one of the “rich response” tools described above

The result was a set of 12 “clusters” of tasks. Sometimes the 3 variants were, superficially, almost identical with just the “space for working” missing from variant C1; in other cases, where it seemed clear that the task could not be presented in all three modes without major revision, the differences were more radical.

The complete task set (including working computer versions) and mark schemes can be found on the Appendix CD-ROM.

Structure of the trials

The generic tasks were grouped into two “tests” as shown in Table 6.1. Three versions of each test were then assembled: a conventional paper test and two computer versions containing complementary mixtures of the C1 and C2 task variants.

Test 1						
Paper		Computer A		Computer B		Source
Balls	P1	Balls	C1	Balls	C2	G
Triangle	P1	Triangle	C2	Triangle	C1	G
Currency	P1	Currency	C1	Currency	C2	G
Trip	P1	Trip	C2	Trip	C1	G
Van Hire	P1	Van Hire	C1	Van Hire	C2	B
Sofa	P1	Sofa	C2	Sofa	C2	W
Test 2						
Paper		Computer A		Computer B		Source
Glass	P1	Glass	C2	Glass	C1	B
Concrete	P1	Concrete	C1	Concrete	C2	G
Percentages	P1	Percentages	C2	Percentages	C1	G
Lines	P1	Lines	C1	Lines	C2	G
Eggs	P1	Eggs	C2	Eggs	C1	B
Taxi Times	P1	Taxi Times	C1	Taxi Times	C2	B

Sources G: original based on a GCSE pattern;
 B: Adapted from MARS *Balanced Assessment* tests;
 W: *World Class Tests*

Table 6.1: The six test variants used in the trials

The trial was based on a cross-over model: each student would take the paper version of one of the two tests and either the “A” or “B” version of the other test on computer.

The aim was to compare, as independent samples, the score distributions of students taking each variant of a task. This avoids the problem (as seen in the *Progress in Maths* data) of the same students taking two variants of the same question within a short (and uncontrollable) period of time.

Ideally, each student would have been randomly assigned a permutation of paper and computer tests. However, while randomising the computer tests could easily be organised by the on-line registration system, having two different paper tests in one school and ensuring each student was handed the correct paper would have been complicated and error-prone. Consequently, each student in a particular school took the same paper test but was randomly assigned either the “A” or “B” version of the complementary computer test.

When a school agreed to take the test, the teacher was given an account on the online registration system and invited to register the pupils to take the test. During this process, teachers were also asked to enter the name, age, gender, Key Stage 3 Maths and English levels and predicted GCSE maths grade for each student. The computer then assigned each student an ID number and a random password.

6 - Design and evaluation of a prototype eAssessment system

Teachers were then given the opportunity to print out a list linking ID numbers, names and passwords for their own use. After this, the pupils' real names were discarded from the database, so all access by researchers was anonymous, by ID number.

Marking

Many of the questions could be marked automatically by the system. However, in this study the tasks were also marked “manually” by experienced GCSE markers, who also commented on, and helped to refine, the mark schemes.

Where questions were closely based on actual GCSE tasks, the GCSE mark schemes were used as the starting point. Some had to be re-written slightly to make them easier to implement algorithmically, in which case the human markers were asked to apply the same rules as the computer.

Although the trial included some experimentation with computer marking algorithms, and the system allowed quite sophisticated rules to be written, this aspect of the trial was treated as a “proof of concept” rather than a quantitative exercise. It is assumed that, provided responses can be captured in a consistent, computer-readable form, marking algorithms of arbitrary sophistication could be developed and tested against substantial banks of sample answers. Here, the interest is in how the mechanism of capturing those responses affects the candidates' performance on the tasks.

Data analysis:

The questions for data analysis were:

- How does the presentation affect the performance of the tasks, and does this match the predicted effects?
- Could the additional data captured by Variants P1 (paper) and C2 (rich computer) have been accurately inferred from Variant C1, or does it represent valuable extra evidence?
- How reliable is “human” marking of variants P1 and C2 compared with automatic marking (which may include some heuristic elements and inferences) of Variant C2?

Qualitative analyses:

Other questions to be considered were:

- What has been learnt about the practicalities of delivering this style of test over the internet?

6 - Design and evaluation of a prototype eAssessment system

- How do students react to computer-based maths tests at this level? To this end, the test ended by asking the student to type in their opinions on the test.
- How can existing task types be adapted to computer delivery and marking without corrupting the assessment goals?
- What insights does the adaptation process give into the quality and validity of existing GCSE task types and examination practices?
- Can such systems be used to deliver new types of task?

6.6: Sample size and composition

Two rounds of school trials were conducted. Schools were asked to enter GCSE students from years 10 or 11 and told that the questions would be similar to intermediate tier GCSE.

One problem was a high drop-out rate for participating schools. While a low response rate to the initial invitations was to be expected, a number of schools who agreed to participate subsequently withdrew.

For the first round, 40 schools in the Nottingham/Derby area were invited, of whom 13 responded and were registered on the system. At one stage over 400 students were expected to take part. However, several schools dropped out before taking the tests and the final tally of successfully completed and marked computer tests was 159 students from 4 schools.

A second round of trials produced a response from 7 schools with 364 promised students, which reduced to 142 students from 5 schools after drop-outs.

The final numbers of participating students are shown in Table 6.2 - note that each student took one of the two paper tests and either the A or B variant of the complementary computer test. Discrepancies in the totals for each type of test are due to lost papers, absences, computer failures etc.

	Paper 1	Paper 2	Total Paper	Computer 1A	Computer 1B	Computer 2A	Computer 2B	Total Computer
First Trial	88	86	174	44	45	37	33	159
Second Trial	96	36	132	25	21	46	50	142
Total	184	122	306	69	66	83	83	301

Table 6.2: Total numbers of successfully completed & marked school trials

As far as can be determined, this attrition was due to time pressures, availability of facilities, failure of the school internet connection and other mishaps (including a school burning down and, in one unfortunate case, an OFSTED inspection) rather than fundamental problems with the system.

6 - Design and evaluation of a prototype eAssessment system

Since schools were allocated to one of the two paper tests in advance, so that papers could be printed and posted, each drop-out removed a large tranche of students from one arm of the cross-over. Also, since the ability range of students was often correlated with the school, or the particular class taken by the participating teacher, the low number of schools made it difficult to get comparable samples of students with regard to ability.

Table 6.3 summarises the gender, age, Key Stage 3 attainment levels and teachers' predictions of GCSE grades for the students who successfully attempted both a paper test and a computer test. It can be seen from these charts that while the overall sample represents a good spread of attainment, the group taking Paper Test 1 includes a disproportionate number of younger, high attainers compared to the Paper Test 2 group.

In summary, although respectable numbers of students took each test to produce useful data on the difficulty of individual tasks, caution should be exercised when making comparisons between task variants, as the samples may not be comparable in terms of ability. The system allows results to be filtered according to ability so that more comparable populations can be compared, but this results in quite low numbers.

However, since the A or B versions of the computer test were assigned randomly within classes, comparisons between C1 and C2 task variants usually involve comparable ability ranges and school environments, and may be more informative.

6 - Design and evaluation of a prototype eAssessment system

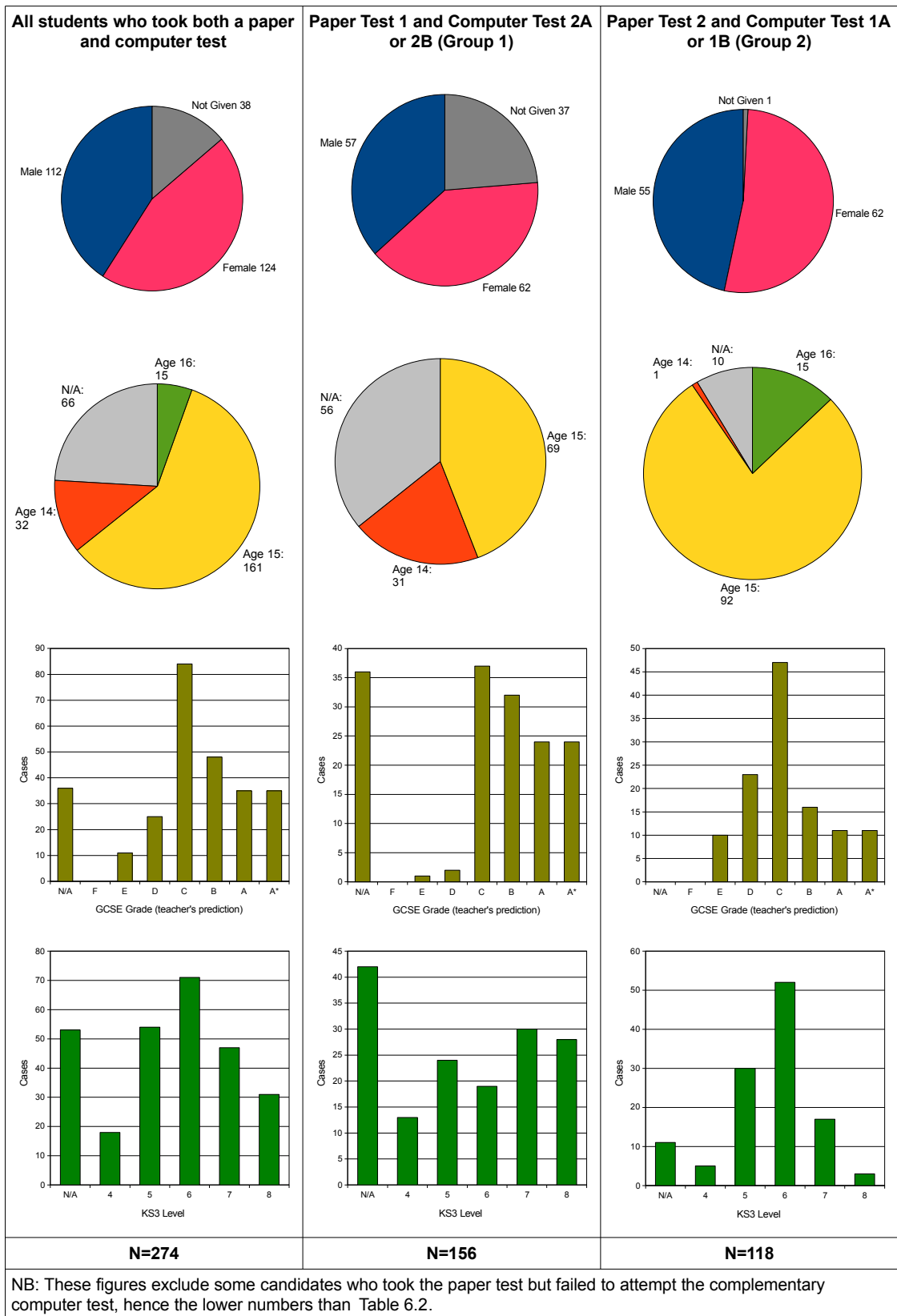


Table 6.3: Students who took both a paper and computer test - composition of sample

6.7: Results – some case studies

Table 6.4 shows summary statistics for all 12 tasks, aggregated over the two rounds of trials. The differences between the paper and computer responses can be largely attributed to the difference in ability levels between the two groups (see table 6.3). However, since the C1 and C2 computer versions were randomly assigned within the same classes, differences between these are potentially interesting. These statistics include a chi-square test comparing the score frequency distributions (where appropriate) and also the percentage of the sample gaining full marks (which is helpful for those task variants with different numbers of points).

It should be noted that chi-squared can produce inaccurate results with small samples, especially if the “expected numbers”, which appear as a divisor in the formula, are less than about five³³. This could have been partially addressed here by pooling the scores into just “pass/fail” rather than looking at each possible score distribution – although this would partly depend on making a task-by-task judgement of what score constituted a pass. This was not done here because of the other known limitations of the data.

The results for just those candidates reported as being at Key Stage 3 levels 4-6 should be more informative – table 6.5 shows summary statistics for that group. Even within this subset it should be noted that group 2 is skewed towards level 6 (as can be seen from the bottom row of Table 6.3) and also that the numbers taking each computer variant are rather low.

It is also informative to consider how the performance on each individual sub-task or mark scheme point varies, particularly in those tasks where some of the variants had extra or different sub-tasks. This is also useful when considering the role of the method part-marks on the paper and C2 versions, and the efficacy of the automatic marking. This data is summarised in table 6.7.

Table 6.6 reviews the discrepancies between the results from the human markers and those from experimental computer marking. These were, subjectively, grouped into three classes:

- **Human error:** the mark given by the human marker is clearly indefensible in terms of the question.
- **Human judgement:** the mark given by the human marker is defensible, but not permitted by a strict interpretation of the mark scheme such as might be made by a computer. For example, a correct result may have been obtained by an obviously wrong method. In a real examination, such issues would probably be raised during the moderation process.
- **Computer error:** the automatic mark is clearly indefensible, suggesting an error, omission or weakness in the computer marking algorithm

³³ leading to $(O-E)^2/E$ becoming very large, where O is the observed number and E is the expected number

6 - Design and evaluation of a prototype eAssessment system

For each task, the table shows the percentage of individual mark scheme “points” affected, and the percentage of candidates whose score on the task would be affected.

Note that the label “human error” as applied here includes mis-marks arising from errors, ambiguities and other design weaknesses in the mark schemes, and is not necessarily a mistake by the marker. In the same way “computer error” is (entirely) attributable to inadequacies in the marking algorithm. Another round or two of revisions and re-trials of both the “human” mark schemes and the computer marking techniques, along with a more rigorous checking and moderation process for the human marking, would be required before making any serious quantitative comparisons. At this prototypical stage, the aim is to gather “proof-of-concept” evidence that the system captures responses in a markable form.

However, human markers are not perfect, and the number of issues with the human marking highlighted by these discrepancies shows that the computer marking can, at a minimum, make a valuable contribution as a method of checking human marking.

Task	Group		Paper (P1)					Simple computer (C1)					Rich computer (C2)					χ ² test: p(null)			Full marks			
	Paper	Comp	N	Out of	Mean	SD	Median	N	Out of	Mean	SD	Median	N	Out of	mean	SD	Median	C1 v P	C2 v P	C2 v C1	P1	C1	C2	
Paper 1																								
Triangle	G1	G2	182	5	51%	42%	60%	66	5	22%	31%	0%	69	5	47%	38%	60%	0.00	0.00	0.00	25%	0%	4%	
Van hire	G1	G2	182	4	58%	38%	50%	64	4	37%	36%	25%	65	4	35%	37%	25%	0.00	0.00	0.01	37%	16%	14%	
Currency	G1	G2	182	4	69%	30%	75%	65	4	58%	28%	50%	69	4	64%	28%	50%	0.01	0.05	0.04	38%	18%	29%	
Balls	G1	G2	184	3	71%	40%	100%	69	3	52%	50%	100%	66	3	51%	48%	33%	n/a	0.00	n/a	62%	52%	47%	
Trip	G1	G2	182	6	75%	22%	83%	67	5	85%	23%	100%	65	6	73%	20%	83%	n/a	0.12	n/a	25%	55%	12%	
Sofa	G1	G2	182	5	33%	41%	20%						124	5	15%	28%	0%	n/a	0.00	n/a	24%		6%	
Paper 2																								
Lines	G2	G1	122	5	23%	30%	10%	82	4	52%	36%	50%	82	5	55%	40%	60%	n/a	0.00	n/a	3%	22%	29%	
Glass	G2	G1	122	4	7%	23%	0%	83	4	20%	39%	0%	83	4	27%	43%	0%	0.00	0.00	0.08	4%	16%	24%	
Concrete	G2	G1	122	2	64%	47%	100%	83	2	66%	48%	100%	82	2	71%	45%	100%	0.07	0.19	0.21	61%	66%	70%	
Taxi*	G2	G1	122	4	17%	31%	0%						157	4	36%	39%	25%	n/a	0.00	n/a	10%		20%	
Percentages	G2	G1	122	7	31%	36%	0%	50	4	40%	36%	50%	46	7	41%	36%	43%	n/a	0.00	n/a	11%	18%	11%	
Percentages†	G2	G1						32	4	53%	35%	0%	37	7	45%	34%	0%	n/a	0.00	n/a		28%	14%	
Eggs‡	G2	G1	122	3	61%	32%	67%	81	1	84%	37%	100%	80	4	63%	24%	75%	n/a	n/a	n/a	30%	84%	4%	

* Ignoring never-awarded 5th point on paper version

† Version with error

‡ Three different tasks

Table 6.4: Summary results - whole sample

6 - Design and evaluation of a prototype eAssessment system

Task	Paper (P1)					Simple computer (C1)					Rich computer (C2)					χ ² test: p(null)			Full marks		
	N	Out of	Mean	SD	Median	N	Out of	Mean	SD	Median	N	Out of	Mean	SD	Median	C1 v P	C2 v P	C2 v C1	P1	C1	C2
Paper 1																					
Triangle	77	5	20%	33%	0%	47	5	23%	33%	0%	45	5	51%	38%	80%	0.13	0.00	0.00	5%	0%	7%
Van hire	77	4	39%	37%	25%	45	4	33%	34%	25%	41	4	28%	36%	0%	0.20	0.20	0.02	16%	13%	12%
Currency	77	4	56%	31%	50%	46	4	54%	26%	50%	45	4	61%	29%	0%	0.05	0.35	0.32	22%	11%	24%
Balls	77	3	55%	43%	67%	45	3	42%	50%	0%	47	3	49%	48%	33%	n/a	0.02	n/a	39%	42%	45%
Trip	77	6	62%	21%	67%	43	5	80%	26%	80%	46	6	74%	21%	83%	n/a	0.00	n/a	6%	42%	17%
Sofa	77	5	9%	24%	0%	0	0				81	5	9%	20%	0%	n/a	0.34	n/a	5%		1%
Paper 2																					
Lines	89	5	17%	24%	0%	31	4	25%	29%	0%	26	5	27%	28%	20%	n/a	0.01	n/a	0%	0%	0%
Glass	89	4	5%	16%	0%	26	4	1%	5%	0%	32	4	2%	9%	0%	0.43	0.83	0.94	1%	0%	0%
Concrete	89	2	65%	46%	100%	32	2	53%	51%	100%	26	2	46%	51%	0%	0.03	0.03	0.73	62%	53%	46%
Taxi*	89	4	14%	24%	0%						50	4	16%	29%	0%	n/a	0.56	n/a	0%		8%
Percentages	89	7	24%	30%	0%	7	4	11%	20%	0%	11	7	10%	23%	0%	n/a	0.71	n/a	4%	0%	0%
Percentages†						19	4	34%	28%	50%	21	7	27%	25%	43%	n/a	0.01	n/a	4%	5%	0%
Eggs‡	89	3	58%	31%	67%	25	1	68%	48%	100%	29	4	52%	31%	75%	n/a	n/a	n/a	25%	68%	0%

* Ignoring never-awarded 5th point on paper version

† Version with error

‡ Three different tasks

Table 6.5: Summary results – pupils at KS3 maths levels 4-6

	Mis-marked points			Mis-marked candidates			Comments			
	Human Error	Human Judgement	Computer Error	Human Error	Human Judgement	Computer Error				
Triangle C1	5.6%	0.0%	0.5%	6	0	1	9.1%	0.0%	1.5%	Mostly awarding 1 points for parts worth 2
Triangle C2	4.3%	0.6%	0.6%	7	2	3	10.1%	2.9%	4.3%	Human errors (first question?); Text mismatches
Van Hire C1	0.8%	0.0%	0.0%	2	0	0	3.1%	0.0%	0.0%	
Van Hire C2	4.2%	3.8%	0.8%	10	10	1	15.4%	15.4%	1.5%	Various issues
Currency Exchange C1	0.8%	0.0%	0.0%	1	0	0	1.5%	0.0%	0.0%	Human error
Currency Exchange C2	0.4%	2.5%	0.0%	1	7	0	1.4%	10.1%	0.0%	Revise mark scheme – is \$569.1 acceptable?
Balls C1	0.0%	0.0%	0.0%	0	0	0	0.0%	0.0%	0.0%	
Balls C2	3.0%	2.5%	0.0%	6	5	0	9.1%	7.6%	0.0%	Unclear/incorrect mark scheme; Missing follow-through
Trip C1	1.5%	0.0%	3.7%	4	0	10	6.0%	0.0%	14.9%	Text matching needs refinement – no negatives to test against
Trip C2	1.0%	0.5%	2.3%	4	2	9	6.2%	3.1%	13.8%	Text matching needs refinement – insufficient negatives to test against
Lines C1	0.8%	2.8%	0.0%	2	7	0	2.4%	8.5%	0.0%	Markers more strict about ambiguous label placement
Lines C2	4.9%	2.0%	1.2%	10	5	3	12.2%	6.1%	3.7%	Inaccurate human marking; some borderline answers
Glass C1	1.8%	0.0%	0.0%	2	0	0	2.4%	0.0%	0.0%	
Glass C2	0.6%	1.2%	0.0%	1	2	0	1.2%	2.4%	0.0%	
Concrete C1	0.0%	0.0%	0.0%	0	0	0	0.0%	0.0%	0.0%	
Concrete C2	0.0%	1.2%	0.0%	0	1	0	0.0%	1.2%	0.0%	
Percentages C1 (with error)	1.0%	0.0%	0.0%	1	0	0	3.1%	0.0%	0.0%	
Percentages C1 (2)	0.0%	0.0%	0.0%	0	0	0	0.0%	0.0%	0.0%	
Percentages C2 (with error)	3.9%	0.0%	0.4%	8	0	1	21.6%	0.0%	2.7%	Markers misapplying mark scheme
Percentages C2 (2)	2.2%	4.3%	0.0%	7	5	0	15.2%	10.9%	0.0%	Markers being (over) generous with unclear marking
Eggs C1	1.2%	0.0%	0.0%	1	0	0	1.2%	0.0%	0.0%	
Eggs C2	0.6%	0.0%	1.9%	2	0	6	2.5%	0.0%	7.5%	Borderline answers – refine tolerances and markers' overlay
Taxi Times	3.3%	5.7%	4.0%	19	34	24	12.1%	21.7%	15.3%	Both computer and human mark schemes need refinement; pupils poor at showing working clearly

Table 6.6: Computer vs. human marking discrepancies - summary

6 - Design and evaluation of a prototype eAssessment system

Whole sample

Key Stage 3 Levels 4-6 only

Task	Part	Paper	C1		C2	
			Human	Auto	Human	Auto
Triangle	1.1	(a) method	62%		62%	58%
	1.2	(a) answer	59%	29% 38%	61%	58%
	2.1	(b) method	53%		51%	42%
	2.2	(b) answer	51%	17% 24%	52%	42%
	3	(c)	31%	0% 2%	7%	6%
Van Hire	1	(a) answer	59%	33% 33%	35%	34%
		(b) method	59%	53% 56%	42%	40%
		(b) method	71%	36% 36%	37%	32%
	2	(b) complete	56%	27% 27%	26%	25%
Lines	1.1	(a) $y=2$	34%	67% 72%	67%	65%
	1.2	(a) $y=2x-1$	11%	44% 44%	40%	46%
	2	(b) intersection	33%	52% 50%	57%	57%
Sofa	1		44%		26%	
	2	Part marks	33%		13%	
	3		33%		15%	
	4		30%		15%	
	5	Fully correct	24%		6%	
Glass	1		11%	22% 23%		
	2	Incompatible mark schemes	10%		30%	30%
	3		5%			
	4		4%	16% 17%	24%	23%
Currency	1.1	(a) method	91%		90%	90%
	1.2	(a) answer	88%	77% 75%	87%	87%
	2.1	(b) method	56%		45%	48%
	2.2	(b) answer	39%	29% 29%	33%	25%
	Concrete	1.1	Method	66%		72%
1.2		Answer	61%	66% 66%	70%	70%
Balls	1.1	Method	82%		50%	59%
	1.2	Method	72%		55%	47%
	1.3	Answer	62%	52% 52%	47%	47%
Trip	1	(a) answer	97%	97% 97%	98%	94%
	2	(b) answer	93%	93% 91%	94%	94%
	3.1	(c) method	85%		78%	78%
	3.2	(c) answer	64%	85% 85%	77%	75%
	4	Graph	98%		23%	20%
Taxi Times	5	Meeting point	94%	64% 63%	69%	71%
	1	Fit line	19%		51%	65%
	2	Av. speed calc.	10%		29%	1%
	3	Av. speed correct	22%		27%	26%
	4	Answer	19%		36%	39%
Percentages	5	Comment	0%			
	1.1	(a) method	49%		57%	46%
	1.2	(a) Tara calc.	41%	58% 58%	43%	35%
	1.3	(a) Julie calc.	43%	60% 58%	52%	43%
	1.4	(a) correct	39%		46%	28%
	2.1	(b) method	17%		30%	30%
	2.2	(b) method	15%		30%	30%
Percentages Error in (b)	2.3	(b) correct	12%	20% 20%	30%	30%
	1.1	(a) method			73%	73%
	1.2	(a) Tara calc.		75% 75%	57%	54%
	1.3	(a) Julie calc.		81% 81%	65%	65%
	1.4	(a) correct			57%	38%
	2.1	(b) method			22%	24%
	2.2	(b) method			22%	24%
Eggs	2.3	(b) correct		28% 28%	22%	24%
		Best fit	80%		84%	83%
		Describe rel.	41%	84% 85%		
		Spot error 1			8%	8%
		Spot error 2			88%	84%
Triangle		Predict length	64%		74%	74%
	1.1	(a) method	30%		67%	62%
	1.2	(a) answer	27%	34% 36%	67%	62%
	2.1	(b) method	17%		56%	42%
	2.2	(b) answer	17%	21% 21%	58%	42%
	3	(c)	9%	0% 2%	7%	7%
	Van Hire	1	(a) answer	38%	31% 31%	32%
		(b) method	38%	51% 56%	29%	34%
		(b) method	55%	29% 29%	29%	27%
2		(b) complete	38%	20% 20%	22%	20%
Lines	1.1	(a) $y=2$	26%	42% 55%	50%	46%
	1.2	(a) $y=2x-1$	6%	16% 16%	8%	12%
	2	(b) intersection	30%	26% 26%	31%	23%
Sofa	1		19%		21%	
	2	Part marks	8%		6%	
	3		6%		7%	
	4		8%		11%	
	5	Fully correct	5%		1%	
Glass	1		9%	0% 0%		
	2	Incompatible mark schemes	8%		3%	3%
	3		1%			
	4		1%	0% 0%	0%	0%
Currency	1.1	(a) method	86%		87%	87%
	1.2	(a) answer	79%	78% 76%	84%	84%
	2.1	(b) method	38%		44%	49%
	2.2	(b) answer	22%	22% 22%	29%	22%
Concrete	1.1	Method	69%		46%	50%
	1.2	Answer	62%	53% 53%	46%	46%
Balls	1.1	Method	69%		47%	60%
	1.2	Method	56%		55%	45%
	1.3	Answer	39%	42% 42%	45%	45%
Trip	1	(a) answer	95%	95% 95%	98%	93%
	2	(b) answer	93%	93% 91%	98%	98%
	3.1	(c) method	81%		76%	76%
	3.2	(c) answer	49%	81% 81%	74%	74%
	4	Graph	98%		30%	26%
Taxi Times	5	Meeting point	98%	49% 47%	65%	67%
	1	Fit line	15%		26%	40%
	2	Av. speed calc.	3%		10%	0%
	3	Av. speed correct	20%		8%	10%
	4	Answer	16%		20%	22%
Percentages	5	Comment	0%			
	1.1	(a) method	44%		18%	18%
	1.2	(a) Tara calc.	34%	14% 14%	18%	18%
	1.3	(a) Julie calc.	35%	29% 29%	18%	18%
	1.4	(a) correct	31%		18%	18%
	2.1	(b) method	10%		0%	0%
	2.2	(b) method	8%		0%	0%
Percentages Error in (b)	2.3	(b) correct	6%	0% 0%	0%	0%
	1.1	(a) method			57%	57%
	1.2	(a) Tara calc.		58% 58%	43%	38%
	1.3	(a) Julie calc.		68% 68%	48%	48%
	1.4	(a) correct			43%	24%
	2.1	(b) method			0%	0%
	2.2	(b) method			0%	0%
Eggs	2.3	(b) correct		5% 5%	0%	0%
		Best fit	81%		72%	72%
		Describe rel.	34%	68% 68%		
		Spot error 1			3%	3%
		Spot error 2			72%	69%
	Predict length	60%		59%	59%	

Table 6.7: Facility levels (average % score) on individual mark scheme points

6.8: Detailed results for selected tasks

The shortcomings of the data discussed above make it difficult to confirm or refute the existence of any systemic difficulty between the various modes of presentation.

Rather than produce a long commentary on largely inconclusive data, here we will look in detail at a small number of tasks for which the results did appear to suggest an effect. Given the low numbers involved, however, all the observations in the following section should be read as suggestions for future investigation rather than conclusive findings.

Triangle

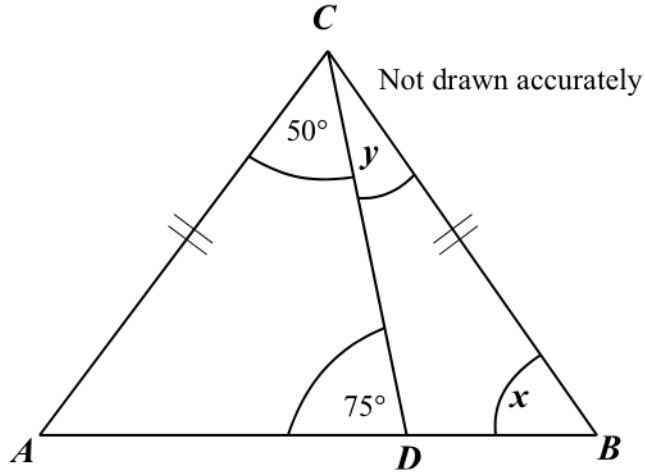
The task represents a common genre of GCSE question, for example, AQA, 2003 question 4. Here, although a different triangle and different missing angles have been used, the structure of the question is the same as the GCSE version, as are the required skills of deducing missing angles in a triangle by identifying equal angles and using the knowledge that all the internal angles add up to 180° . Figure 6.9 Shows the paper task used in the trials, which takes the typical GCSE pattern of giving spaces for final, numerical answers preceded by an area for showing working. Figure 6.10 shows the “rich” computer version (C2), which uses the “printing calculator” tool discussed above in place of the space for working.

The intermediate computer version, not shown, was very similar to the C2 version, but without the spaces for working. In all cases, there was a final part to the question requiring a short, text answer saying whether or not the final assertion was true.

The inclusion of this task was partly intended to explore the efficacy of using candidates' working to award partial credit, particularly for fairly simple computations with short “reasoning length” and only one intuitive step. Accordingly, the mark scheme (Figure 6.11) followed the pattern of the GCSE original, with two marks for a correct final answer, or a single mark if the correct working is seen³⁴. If this was benefiting candidates who made a mistake, but showed evidence of correct working, it would be expected that the proportion of candidates receiving the method mark would be significantly higher than those receiving both.

³⁴ Based on AQA's “Notes for Examiners” (AQA, 2003) – confirmed by the experienced GCSE markers who worked on the trials - it is GCSE convention to award full marks (i.e. both the “A” and “M” marks in this case) for a correct answer, unless the question specifically asked for working to be shown.

ABC is a triangle.
 $CA=CB$.
 D is a point on AB .



- (a) Work out the value of x .

.....

Answer $x=.....$ degrees (2 marks)

- (b) Work out the value of y .

.....

Answer $y=.....$ degrees (2 marks)

- (c) Does $AD=DC$?
 Give a reason for your answer.

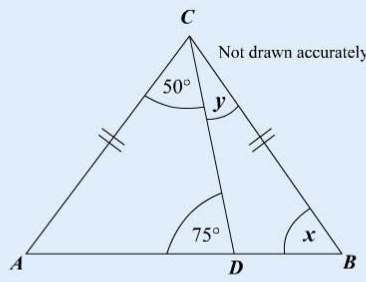
.....

(1 mark)

Figure 6.9: Triangle task (paper version)

Triangle

ABC is a triangle.
CA = CB.
D is a point on **AB**



Not drawn accurately

Show how you worked out your answers:

Use the on-screen calculator and drag the "printout" to your answer.

(a) Work out the value of x .

Drop the printout from the calculator here to show your working

Answer: $x =$ degrees

(2 marks)

(b) Work out the value of y .

Drop the printout from the calculator here to show your working

Answer: $y =$ degrees

(2 marks)

Task 1 of 11

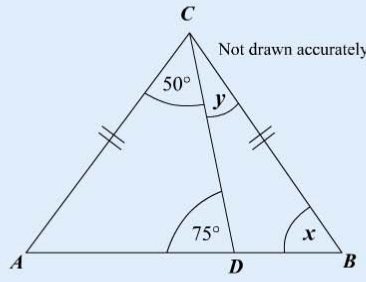
Page 1 of 22

Next

Quit

Triangle

ABC is a triangle.
CA = CB.
D is a point on **AB**



Not drawn accurately

Show how you worked out your answers:

Use the on-screen calculator and drag the "printout" to your answer.

(c) Does **AD=DC**?
 Give a reason for your answer.

Answer:

(1 mark)

Task 1 of 11

Back

Page 2 of 22

Next

Quit

Figure 6.10: Triangle task (C2 variant - computer with rich responses)

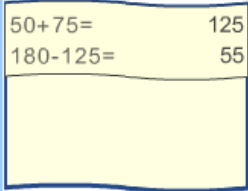
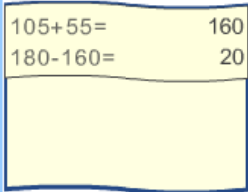
2) Triangle		Attempt 1 of 1	
Markscheme		Response	C M
1.1	180-75-50 seen	M1 	1 1
1.2	55	A1 Answer= 55	1 1
2.1	180-55-55-50 seen	M1 	1 1
2.2	20	A1 Answer= 20	1 1
3	No because 75 and 55 are not equal	B1 no AD- 130 DC-95	0 0

Figure 6.11: Triangles mark scheme and sample response (C2 variant)

- (a) Work out the value of
- x
- .

$$\Delta = 180 \quad \angle CAD = 55^\circ \quad x = \angle CAD \text{ (isosceles)}$$

$$180 - (75 + 50) = 55$$

Answer $x = 55^\circ$ degrees (2 marks)

Figure 6.12: Fully justified answer to part (a) of triangle - from trials of paper test

One possible criticism of the printing calculator is that students can only present calculations as working – they can not make a complete argument such as that shown in Figure 6.12.

However, that response exceeds what is usually expected for this type of question: the mark-schemes used here, which followed the pattern of those from the source GCSE questions, offered no credit for explanation or statements of assumptions – just full credit for the correct answer and the chance of partial credit where computations such as “180 – 50 – 75” were seen.

Out of all the 107 **correct** responses to this part of the paper test:

- 83% provided working (usually showing the calculations)
- under 5% (5 out of 107) also gave explanations comparable to Figure 6.12 (the most complete explanation seen)

6 - Design and evaluation of a prototype eAssessment system

- 25% gave or implied some justification (such as saying that $A=B$ or $A=55$; $x=55$ but not saying why)
- 36% implied justification by annotating the diagram – usually by labelling $\angle CAD$ as 55 or x .

From the 53 **incorrect but non-blank** responses:

- 53% provided working (usually showing the calculations)
- 9% (5 out of 53 – or less than 3% of the total sample) benefited from part marks for working
- 1 indicated that $\angle CAD=x$ in their working
- 1 implied this on the diagram
- 15% (8 out of 53) made similar, but wrong comments or annotations

So, in summary, the majority who got the answer wrong also failed to provide correct working, and even if the mark scheme had offered partial credit for (say) indicating that $\angle CAD=x$ only a few candidates would have benefited. This suggests that the ability to provide working in the computer version would not have a significant influence on the scores. The ability to annotate the graph would be equally useful.

However, there is the (unsurprising) suggestion that those pupils who showed good reasoning were also the ones who got the answer correct: this raises the question of whether removing the ability to gain part marks from working would have a negative effect on scores.

Table 6.8 shows the percentage of trial candidates receiving each point in the mark scheme for each variant of the task. It can be seen that the facilities for the “Method” and “Answer” points are very similar, especially considering that a few mis-marked scripts are to be expected. As noted earlier, this suggests that the method marks are not making any obvious, significant contribution to the psychometrics of this task: this is hardly surprising in a question of this type where students are using calculators, making it very likely that a candidate who had identified the correct calculation would proceed to the correct answer.

6 - Design and evaluation of a prototype eAssessment system

Whole Sample	N	% Facility for (a)		% Facility for (b)		% Facility for (c)
		Method	Answer	Method	Answer	Answer
Paper	182	62	59	53	51	31
Computer C1	66		38 ³⁵		24	0
Computer C2	69	62	61	51	52 ³⁶	7
KS3 Level 4-6						
Paper	77	30	27	17	16	9
Computer C1	47		36		21	0
Computer C2	45	67	67	56	58	

Table 6.8: Facilities for the individual parts of the "Triangle" task (original marking)

It can be seen that, although the awarding of method marks had little effect, the C1 version of the task still proved substantially more difficult than the C2 version (compare the "Answer" facility levels in Table 6.8).

The presentation of the C1 and C2 versions were visually very similar and both included a highly visible on-screen calculator. The only obvious difference is that the C2 version included the "printing calculator" functionality, and pupils were explicitly instructed to use this to show working. This raises the possibility that having to provide working, even if it was not directly credited, could somehow be improving performance on the task.

The results from the paper version are ambiguous due to the limitations of the sample: insofar as the results can be filtered to a comparable ability range, they suggest that performance on the paper task, despite allowing working, was more comparable to the C1 short answer variant.

Table 6.9 shows the percentages of pupils providing working – the differences between paper and computer seem unlikely to be significant and could, for example, be attributed to the higher rate of totally blank responses on paper. These comparisons do show a clear correlation between showing working and getting the correct answer.

Triangle Paper part (a) – all candidates

N=182	Right	Wrong	Blank	Tot.
Working	49%	15%	1%	64%
No working	10%	13%	13%	36%
Total	59%	28%	13%	

Triangle C2 part (a) – all candidates

N=68	Right	Wrong	Blank	Tot.
Working	49%	21%	1%	71%
No Working	9%	12%	9%	29%
Total	57%	32%	10%	

Triangle Paper part (a) KS3 4-6

N=77	Right	Wrong	Blank	Tot.
Working	25%	25%	1%	51%
No working	3%	25%	22%	49%
Total	27%	49%	23%	

Triangle C2 part (a) – KS3 4-6

N=44	Right	Wrong	Blank	Tot.
Working	50%	14%	0%	64%
No Working	11%	14%	11%	36%
Total	61%	27%	11%	

Table 6.9: Percentages of candidates providing working - Triangle Paper vs. C2

35 In this case, the computer-awarded scores were used since a consistent error was spotted in the human marking.

36 This would suggest a marking error since anyone getting the "answer" mark should have got the "method" mark automatically.

6 - Design and evaluation of a prototype eAssessment system

Table 6.10 shows the relative frequencies of answers to the first two parts, and could reveal whether one of the variants was encouraging a particular mistake or misconception. So, for example, the answer (a) $x=60$, (b) $y=10$ arises from assuming the larger triangle is equilateral; while a pupil who answered (a) $x=55$, (b) $y=70$ possibly calculated $\angle ACB$ and forgot to subtract $\angle ACD$. From the paper responses, (a) $x=55$, (b) $y=25$ has been seen to arise from an arithmetic error. In this case, the numbers of students giving the “top” wrong answers are too small to draw conclusions from, but with a larger sample, this sort of analysis could be revealing.

With hindsight, choosing the values of the angles strategically when designing the task could make it easier to distinguish various logical mistakes from arithmetic errors (making all the given angles a multiple of 5 meant that most wrong answers seen were also multiples of 5, making it more likely that the result of an arithmetical error would also match some plausible result of copying, adding or subtracting the given angles).

This question appears to be effective at identifying higher attaining students: on the paper test, the final explanation part was key in identifying candidates working at levels 7 and 8, with the facility for that part rising from negligible at level 6 to 35% at level 7 and 79% at level 8.

Unfortunately, the numbers of level 7 and 8 students taking the computer version were small so it is difficult to draw conclusions about any paper vs. computer differences in the way this part performed. It was surmised that the need to place this final question on a second screen, hiding the candidates' responses to the previous parts, coupled with the cognitive load of having to type in a response might make it harder on screen, and while the data do not contradict this, the numbers are too small to be conclusive.

Note – this section combines information from the original manual marking and data entry exercise (e.g. Table 6.8) with later analyses (tables 6.9-6.10) based on simple computer scoring of the numerical answers. Consequently, a few small discrepancies in facility levels and total numbers are apparent – these are confined to a few percent of cases and are insignificant compared to the effects being considered.

6 - Design and evaluation of a prototype eAssessment system

Triangle P – Whole sample					Triangle C1 – Whole sample					Triangle C2 – Whole sample				
Answer (a)	Answer (b)	Cases	% of all responses	% of wrong responses	Answer (a)	Answer (b)	Cases	% of all	% of wrong	Answer (a)	Answer (b)	Cases	% of all	% of wrong
55	20	88	48%		55	20	16	25%		55	20	28	41%	
Blank	Blank	25	14%	27%	Blank	Blank	5	8%	10%	Blank	Blank	7	10%	18%
50	25	5	3%	5%	55	70	5	8%	10%	60	10	4	6%	10%
55	Blank	4	2%	4%	50	25	5	8%	10%	65	10	3	4%	8%
75	Blank	4	2%	4%	60	10	4	6%	8%	55	30	2	3%	5%
60	10	4	2%	4%	27.5	27.5	3	5%	6%	50	10	2	3%	5%
75	25	3	2%	3%	115	65	2	3%	4%	55	70	2	3%	5%
55	50	3	2%	3%	65	10	2	3%	4%	55	40	2	3%	5%
65	10	2	1%	2%	35	40	2	3%	4%	50	25	2	3%	5%
45	30	2	1%	2%	50	130	2	3%	4%	55	27.5	1	1%	3%
55	25	2	1%	2%	37.5	55	1	2%	2%	105	35	1	1%	3%
55	70	2	1%	2%	50	75	1	2%	2%	65	15	1	1%	3%
40	25	2	1%	2%	65	37.5	1	2%	2%	50	30	1	1%	3%
65	Blank	1	1%	1%	30	25	1	2%	2%	27.5	27.5	1	1%	3%
48	24	1	1%	1%	105	55	1	2%	2%	55	80	1	1%	3%
90	40	1	1%	1%	55	15	1	2%	2%	115	65	1	1%	3%
55	5	1	1%	1%	55	50	1	2%	2%	52.5	10	1	1%	3%
60	20	1	1%	1%	70	30	1	2%	2%	55	62.5	1	1%	3%
45	25	1	1%	1%	80	Blank	1	2%	2%	50	40	1	1%	3%
30	40	1	1%	1%	55	35	1	2%	2%	45	30	1	1%	3%
55	75	1	1%	1%	45	25	1	2%	2%	105	20	1	1%	3%
60	15	1	1%	1%	45	15	1	2%	2%	55	Blank	1	1%	3%
50	Blank	1	1%	1%	50	45	1	2%	2%	52.5	Blank	1	1%	3%
70	20	1	1%	1%	105	75	1	2%	2%	90	90	1	1%	3%
60	30	1	1%	1%	105	130	1	2%	2%	55	10	1	1%	3%
120	20	1	1%	1%	55	100	1	2%	2%					
7.5	7.5	1	1%	1%	105	80	1	2%	2%					
50	21	1	1%	1%	50	40	1	2%	2%					
...and 20 more single cases					50	5	1	2%	2%					
Total		182			Total		65			Total		68		

Table 6.10: Relative frequencies of responses to the Triangle task

Even with the small numbers, there is a strong suggestion that, in this case, something about the C2 variant made the question “easier” than C1 or paper. Since these two were visually very similar, one hypothesis might be that the novelty value of the printing calculator in C2 simply made an otherwise quite dry mathematical exercise more interesting and engaging. This would be consistent with the suggestion that the paper version, even though it allowed working to be shown, was as hard as the C1 version.

Percentages

- (a) Tara earns £260 per week.
She is given a pay rise of 2.5%

Julie earns £220 per week.
She is given a pay rise of 3%

Whose weekly pay increases by the greater amount of money?

.....

Answer: (4 marks)

- (b) Nick has just been given a 4% pay rise.
He now gets £218.92 per week.

How much did he earn each week before the pay rise?

.....

Answer: £..... (3 marks)

Figure 6.13: Percentages - paper version

Adapted from a common GCSE task type (e.g. AQA, 2003 Paper 2 Q12), this task depends more heavily on capturing working than others, such as *Triangle*. Here, the markers must see the pupils' method for part (a) in order to eliminate the possibility that they simply guessed "Julie" was the correct answer.

Like the original paper version (fig. 6.13) the C2 variant asked pupils to say which person would receive the largest pay increase, and used the printing calculator tool to capture the working. However, this was not possible in the simple-answer-only C1 variant. To produce a version without working, it was decided to ask the pupil to calculate the two new salaries, on the assumption that this is how most pupils would tackle the problem anyway (see fig. 6.14). This is clearly no longer the same task, although the check list of mathematical content knowledge required is similar.

<p>(a) Tara earns £260 per week. She is given a pay rise of 2.5%.</p> <p>Julie earns £220 per week. She is given a pay rise of 3%.</p> <p>Whose weekly pay increases by the greater amount of money?</p> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;"> Drop the printout from the calculator here to show your working </div> <p style="text-align: right;">Answer: <input type="button" value="Tara"/> <input type="button" value="Julie"/></p>	<p>(a) Tara earns £260 per week. She is given a pay rise of 2.5%.</p> <p>Julie earns £220 per week. She is given a pay rise of 3%.</p> <p>How much do Tara and July now earn each week?</p> <p>Tara: £ <input type="text"/> per week</p> <p>Julie: £ <input type="text"/> per week</p>
--	--

Figure 6.14: Percentages C2 (left) vs. C1 (right) - note different structure

6 - Design and evaluation of a prototype eAssessment system

A direct adaptation of the original mark scheme for this part of the question is shown in Figure 6.15 - since this is quite complex, an attempt was made to produce a more “algorithmic” version, with a similar effect, which could be reasonably be applied by both the computer and a human marker. The result is shown in Figure 6.16.

Tara: $260 \times \frac{2.5}{100}$ or Julie: $220 \times \frac{3}{100}$	M1	M1 for attempting to calculate either 2.5% of £260 or 3% of £220 M1 for attempting any valid method of comparison i.e. $260 \times \frac{2.5}{n}$ and $220 \times \frac{3}{n}$
(£)6.5(0)	A1	If only both weekly wages are calculated then award 1 mark (special case) but if weekly wages followed by an answer of “Julie” award full marks.
(£)6.6(0)	A1	
Therefore Julie	A1	

Figure 6.15: "Original" GCSE markscheme for part (a) of percentages

It can be seen from tables 6.7 and 6.11 that the computer consistently gave lower marks for part (a) of the question. Examination of the discrepancies showed that this was usually due to markers being over-generous rather than the computer failing to spot correct answers.

Two examples of marker error are shown in Figure 6.16: in the first case, the marker has either failed to spot, or has decided to accept, a factor of 10 error in the calculations (30% and 25% rather than 2% and 3.5%). The computer has not awarded any marks in this case. While awarding 1 or 2 marks for this might be defensible under the original mark scheme, allowing this to “follow through” so that the pupil still gets full marks is clearly over-generous.

The second case is less extreme: 2 marks from the computer versus 3 from the marker. Here, the marker has not followed the given mark scheme strictly: evidence of correct calculations for **both** employees is needed to eliminate the chance of guessing. (The original mark scheme seems to allow 1 mark for the correct name without any other evidence – despite a 50% chance of guessing correctly...)

In this case, it appears that the automatic marking was more consistent, if slightly less generous, than the human markers.

Marking:	5	School:		Start:	11:05	Date:	20 Jun 006
Student ID:	504			End:	11:23	Time:	12:18:20
3) Percentages				Attempt 1 of 1			
Markscheme			Response		C	M	
1.1	6.5 (260 x 2.5/100) or 6.6 (220 x 3/100) seen	M1	220x1.3=	286	0	1	
1.2	6.5(0) seen (or 266.5 seen and answer=Julie)	A1	260x1.25=	325	0	1	
1.3	6.6(0) seen (or 226.6 seen and answer=Julie)	A1	286-220=	66	0	1	
			325-260=	65			
1.4	Julie and some valid working shown (1.2 and 1.3 awarded)	A1	Answer (1=Tara, 2=Julie) 2		0	1	
2.1	1.04 or 104 (%) seen	M1	218.92÷100=	2.1892	0	0	
2.2	218.92 / 104 or 218.92/1.04 seen	M1	Ans x 4=	8.7568	0	0	
			218.92-Ans =	210.1632			
2.3	(218.92 / 1.04=) £ 210.50 Accept 210.5 but not 210.05, 210.5p or 210-5	A1	Answer : £ 210.16 Per week.		0	0	

Marking:	5	School:		Start:	01:41	Date:	11 Jul 006
Student ID:	544			End:	01:58	Time:	12:16:55
3) Percentages				Attempt 1 of 1			
Markscheme			Response		C	M	
1.1	6.5 (260 x 2.5/100) or 6.6 (220 x 3/100) seen	M1	220÷100=	2.2	1	1	
1.2	6.5(0) seen (or 266.5 seen and answer=Julie)	A1	220+ (Ans x 3)=	226.6	0	0	
1.3	6.6(0) seen (or 226.6 seen and answer=Julie)	A1			1	1	
1.4	Julie and some valid working shown (1.2 and 1.3 awarded)	A1	Answer (1=Tara, 2=Julie) 2		0	1	

Figure 6.16: Mark scheme and two examples of marking discrepancies
The "C" column shows the automatic mark, "M" shows the human maker's mark

Trip

The paper variant of this question is shown in Figure 6.17. As with most of the questions in the trial, this was based on a commonly seen type of GCSE question (e.g. AQA, 2003, question 5) with changes to the quantities involved. One change was the addition of the final step (e) for the reason described below.

The rich (C2) computer variant (Figure 6.18) used the drawing tool to allow candidates to draw Anne's distance-time graph in (d) and the printing calculator tool to capture the working for the speed calculation in part (c). This task extended over 5 “pages” with the graph retained in the top 2/3 of the screen and the just current question part changing between screens. In this respect, it differs from the paper versions in that students had to actively move back a page to see their previous answer.

The intermediate (C1) computer version was similar, but did not ask for working in (c) and omitted the graph-drawing step completely – this was the reason for adding the extra “where do they meet” step, as this could be answered by visualising the distance-time graph rather than drawing it.

The mark scheme and sample answer to the C2 variant are shown in Figure 6.19. Table 6.12 summarises the facility levels for individual parts of the lesson.

The GCSE papers examined usually only had one question per paper for which candidates could forfeit a mark by failing to state the correct units. Since having an extra on-screen control on just that question would make this rather obvious³⁸, it was decided to ensure that candidates were asked to state units on several questions even if (as on parts b and c here) these were not actually marked. Ignoring blank/missing responses, all but about 13% of the candidates for the two computer tests supplied correct units.

One possible problem with the task re-design emerged: to make part (e) easy to visualise without drawing the graph, the cross-over point was placed at the midpoint of Anne's distance-time graph. However, this meant that a line drawn sloping in the wrong direction (see the example response in Figure 6.19) still crossed Gavin's graph at the same point. In the trials, this emerged as a common error, so a considerable number of candidates got the correct answer (e) from an incorrect graph.

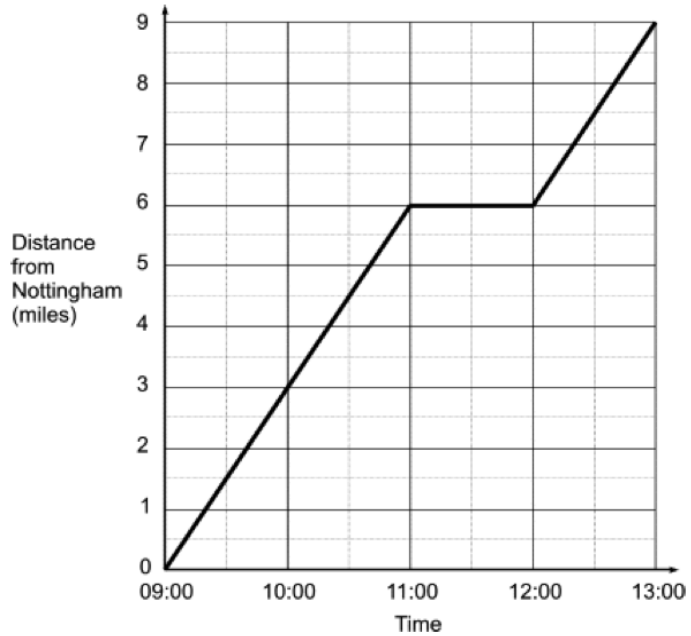
In the context of the paper and C2 versions this makes little difference as, were the symmetry removed, a GCSE-style mark scheme would typically allow “follow through” and credit the correct point of intersection. However, for the C1 variant this means that it could **not** be inferred from (e) that the candidate was visualising the graph correctly.

³⁸ The practise on the GCSE papers examined was, apparently, to have the correct units pre-printed after the space for each answer, with the exception of one question for which a mark was available for writing the correct units.

6 - Design and evaluation of a prototype eAssessment system

Gavin walks from Nottingham to Risley, a distance of 9 miles.

Here is a distance-time graph of his journey:



- (a) What is happening between 11:00 and 12:00?

 (1 mark)
- (b) How far does Gavin travel in the first 2 hours?
 Answer: (1 mark)
- (c) What is Gavin's average speed over the first 2 hours of the journey?

 Answer: (2 marks)
- (d) Anne cycles from Risley to Nottingham.
 She leaves Risley at 10:00 and arrives in Nottingham at 11:00.
 On the diagram, draw a possible distance/time graph of Anne's journey.
 Assume she travels at a constant speed. (1 mark)
- (e) Use the graph to estimate when and where she passes Gavin.
 Time:,
 Distance: miles from Nottingham (1 mark)

Figure 6.17: Trip question - paper version

6 - Design and evaluation of a prototype eAssessment system

Trip

Gavin walks from Nottingham to Risley, a distance of 9 miles.

Here is a distance/time graph for his journey.

(c) What is Gavin's average speed over the first 2 hours?

Drop the printout from the calculator here to show your working

Answer: Units? v

(2 marks)

Show how you worked out your answers:

Use the on-screen calculator and drag the "printout" to your answer.

Task 9 of 12
Back
Page 16 of 23
Next
Quit

Trip

Gavin walks from Nottingham to Risley, a distance of 9 miles.

Here is a distance/time graph for his journey.

Anne cycles from Risley to Nottingham.

She leaves Risley at 10:00 and arrives in Nottingham at 11:00.

(d) On diagram, draw a possible distance/time graph of Anne's journey.

Assume that she travels at a constant speed.

(1 mark)

Plot Points x x

Draw Lines —

Joined-up lines ~

Delete

Task 9 of 12
Back
Page 17 of 23
Next
Quit

Figure 6.18: Trip task (C2 variant)
Parts c (using the printing calculator) and d (using the drawing tool)

6 - Design and evaluation of a prototype eAssessment system

It is clear from the low facility levels for part (d) that the graph-drawing step adds a significant extra demand to the question. Removing this step from the C1 variant has clearly changed difficulty level of this question.

As with “Triangles” the part facilities show that the effect of allowing partial credit for showing the speed calculation was minimal. In the case of the C2 variant, out of 67 candidates, two candidates benefited from partial credit: one because their working showed a correct answer which they had failed to type into the answer box (a mistake they might not have made but for the need to store their working) and the other received a “follow through” mark for correctly dividing their incorrect answer for the distance by 2 hours (which could have been spotted without capturing working).

There is no clear evidence from this question to suggest that having to show working improves performance – if anything, the speed calculation was answered most successfully in the C1 version. It seems unlikely that the fairly trivial sum $(6 \div 2)$ would benefit from the availability of a calculator.

The computer was able to mark this question effectively, with two minor problems:

The case mentioned above, where there was an unambiguous correct answer in the working, was given full credit by the human markers, but part credit by the computer. Rather than having the computer try and pick up such errors at the marking stage where the working might contain other, incorrect calculations and require human judgement to score fairly, it might be better to automatically remind a candidate if they had left the answer box blank.

The simple text-matching used to mark the first part produced false negatives in around 14% of cases - unacceptable for final marking, but accurate enough to be useful for checking human marking. Refining the rules reduced this to about 1.5% with the remaining problems being answers such as “she doesn't walk any more distance, she stays the same and then starts walking again at 12” (correct, but obtuse) or “he only walked 6 miles” (accepted by the human markers, but debatable). Here, the problem was the opposite of that experienced with Triangle – with a facility level of 97-98% on this part there were insufficient test cases to ensure that the system would reliably reject wrong answers.

6 - Design and evaluation of a prototype eAssessment system

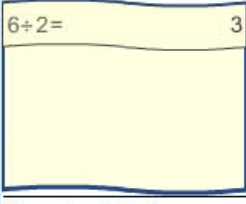
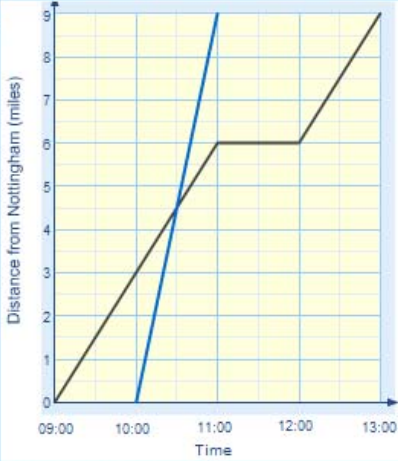
Markscheme		Response	C	M
1	Stopped/not moving/taking a rest/stays same distance	B1 he stops for a rest of some sort.	1	1
2	6 (miles)	B1 Distance= 6 miles	1	1
3.1	(their 6)/2	M1 	1	1
3.2	3 (mph)	A1 Speed = 3 mph	1	1
4	Straight line joining (10:00,9) to (11:00,0)	B1 	0	0
5	Time=10:30; Distance=4.5 (miles)	B1 Time = 10:30 Distance = 4.5	1	1

Figure 6.19: Trip task (C2 variant) mark scheme and sample answer - note that this candidate has drawn the line incorrectly for part 4 (d)

		% Facility					
		(a)	(b)	(c)	(d)	(e)	
All levels		Method		Answer			
Paper	182	96	99	75	74	32	71
Computer C1	67	97	93		85		64
Computer C2	65	98	94	78	77	23	69
KS3 Level 4-6							
Paper	77	95	99	57	56	18	45
Computer C1	43	95	93		81		49
Computer C2	46	98	98	76	74	30	65

Table 6.12: Facilities for the individual parts of the "Trip" task

Taxi Times

As with most of the task set, the two previous questions were derived from GCSE task types, and as such followed the typical pattern of short, well-defined calculations with clear correct answers. *Taxi Times* was, instead, taken from a *Balanced Assessment in Mathematics* task (see section 2.3) which placed more emphasis on the candidate's ability to choose and apply the correct methods. The C2 variant of the task is shown in Figure 6.20. Since this task relies on capturing the graph and working, no C1 variant was produced. Figure 6.21 shows the on-screen mark scheme and a sample answer³⁹ The paper version was similar, but had an extra point available for any sensible comments or caveats about the practicality of the solution (such as allowing for a return trip) – in practice, this was never awarded.

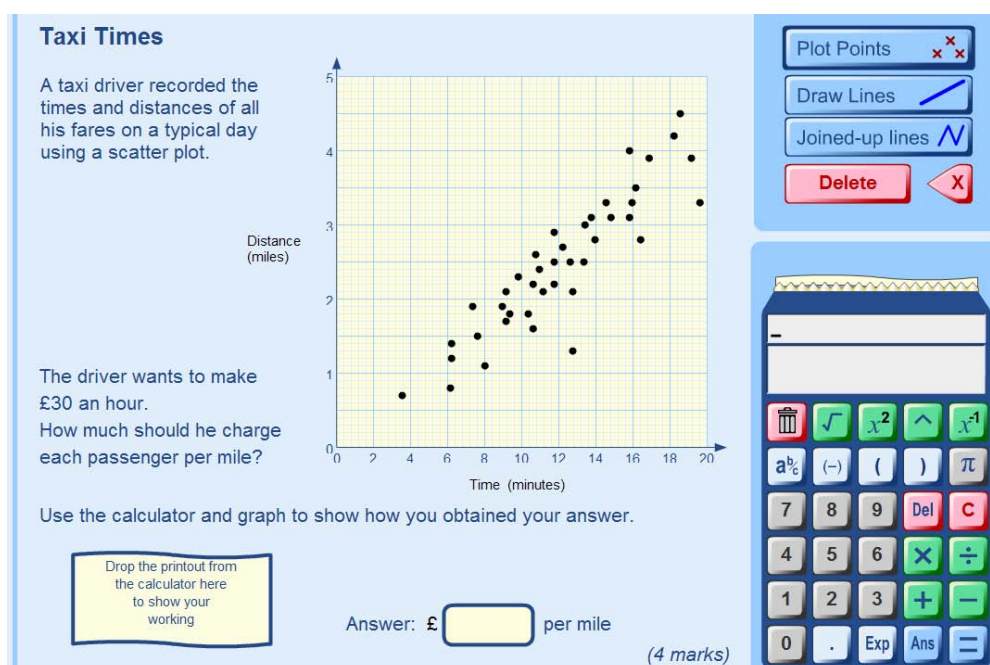


Figure 6.20: Taxi times task (rich computer version)

Figure 6.22 shows a summary of the trial results for this task. It can be seen that only candidates attaining levels 7-8 at Key Stage 3 consistently made substantial progress on this task – although there are a few “outliers” at each level who managed to get good scores.

An interesting observation is that, out of the group that took this question on paper, only 19% successfully drew a line of best fit (most of the rest simply didn't try), yet where an earlier question on the same paper (“Eggs”) specifically asked for a line of best fit on a scatter graph, 80% succeeded. This may be further evidence of how students learn to answer typical GCSE questions without being able to apply the techniques they learn to different types of

³⁹ In the answer shown, the computer disagrees with the human marker over the validity of the calculation, which doesn't match the gradient of the line drawn.

6 - Design and evaluation of a prototype eAssessment system

problems. Anecdotally, one of the experienced GCSE markers employed in the trials, on first seeing this task, commented “this is not on the syllabus” – although questions on either distance/time graphs or drawing lines of best fit on scatter graphs are commonplace.

Performance on the C2 version was noticeably better – but this is probably due to the presence of KS3 level 7 and 8 students in the sample, who were able to make good progress. More students drew lines of best fit in this version – this might be because the presence of the line-drawing tools on-screen provided a strong hint in this direction, not present on paper. It was also apparent that the mark schemes on this type of task – where correct working is required for full marks rather than just as a fallback for partial credit – need careful development and refinement. Discrepancies between human and computer marking revealed overt errors in computer marking in 15% of the candidates' responses – with a larger number of cases in which the human markers had made defensible judgements. However, this also identified about 12% of candidates whose responses were clearly mis-marked by the humans.

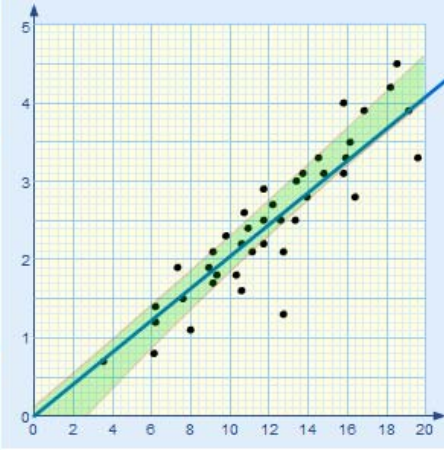
Markscheme		Response	C	M
1	Line of best fit.		1	1
2	Calculates gradient of line or average speed in miles/minute		0	1
3	Average speed between 13 and 16 mph		1	1
4	Result between 1.87 and 2.30	<div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> $60 \div 4.5 = 13.33333333$ $30 \div \text{Ans} = 2.25$ </div> <p>Answer = £ 2.25</p>	1	1

Figure 6.21: Taxi times marking scheme

Apart from some correctable discrepancies in the acceptable tolerances for the line of best fit⁴⁰ a major problem was that rather than simply reading a distance and time from the graph

⁴⁰ The computer's criteria are ranges for y-intercept and gradient, markers were given an overlay with a shaded zone in which the line should fall. This produced some discrepancies in borderline cases.

6 - Design and evaluation of a prototype eAssessment system

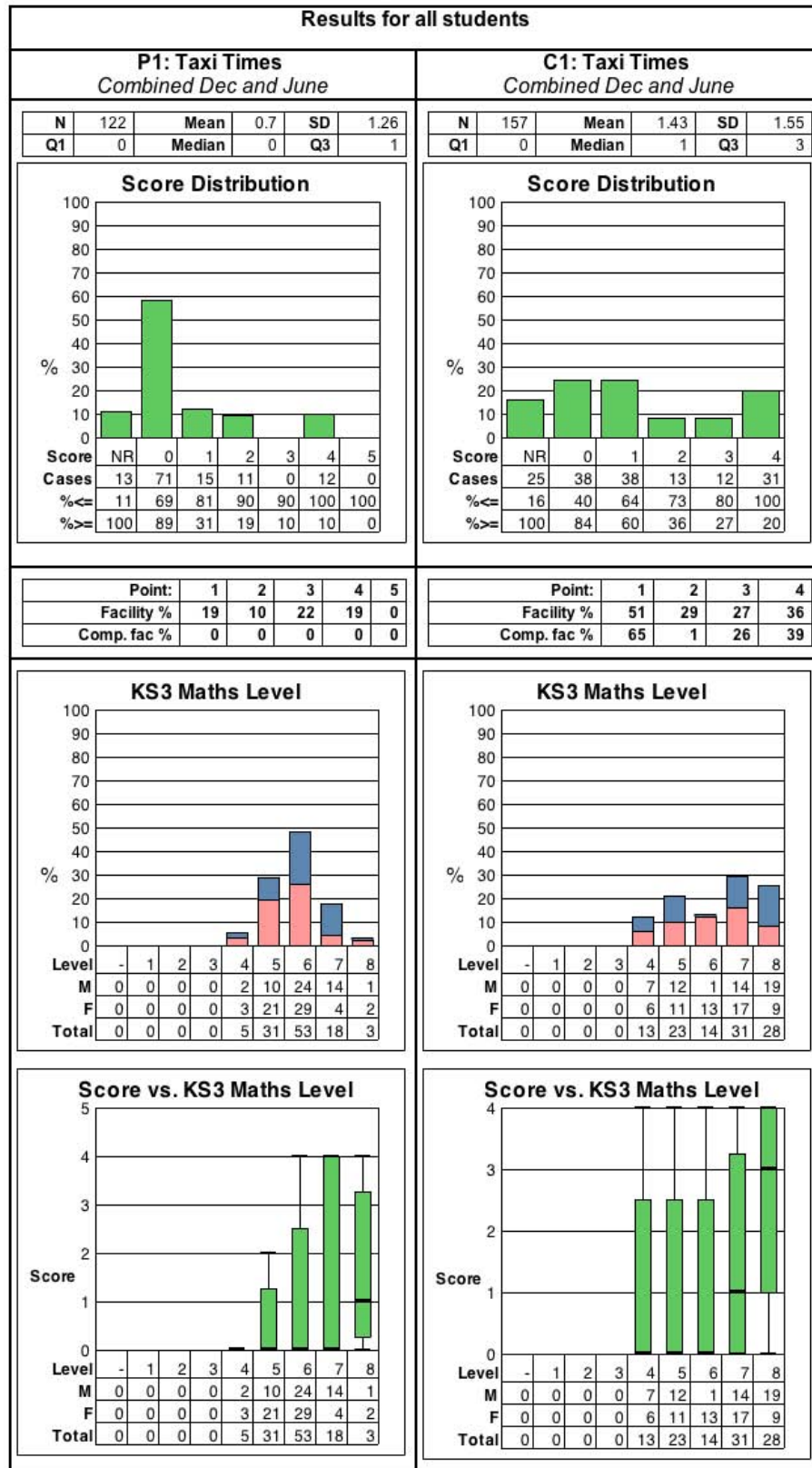


Figure 6.22: Trial results for taxi times task

and dividing on the calculator, then optionally converting to miles per hour, students would often use “shortcut” strategies that combined these steps: such as reading off the distance at 20 minutes and multiplying by 3 to get miles per hour. Although perfectly valid, it would be difficult to identify all these strategies and program the computer to identify them. It is also quite demanding to expect human markers to reliably spot them. For this style of task to become commonplace, candidates would need to be taught that clearly demonstrating their method was an important part of the task.

6.9: Effectiveness of the tools

The printing calculator

The printing calculator tool seemed effective at allowing working to be captured for calculation tasks in a form which either the computer or human markers could interpret.

In some tasks (e.g. *Concrete*, *Triangle*) there is little difference between the facilities for the method mark and the corresponding final answer, suggesting that students either received the method mark by default, having given a correct final answer, or that correct working inevitably led to receiving a correct answer from the calculator. Possibly, this is more a critique on the style of mark schemes which, for these trials, were mostly analogues of schemes from past GCSE papers: if the intention of partial marks is simply to avoid repeatedly penalising poor arithmetic then it is unsurprising that it will serve little purpose where the use of a calculator is encouraged.

In other cases, the method capture is a more vital part of the question: in *Percentages*, *Glass* and *Van Hire* the printing calculator allowed the question to remain substantially as per the paper original (Which is best/which is more/will it fit? - show your working) whereas the C1 “simple answer” version required further scaffolding that explicitly told the pupil what they were expected to calculate. Whether such changes significantly change the difficulty of the question is unclear and depends on the design decisions made when adapting the task: the only obvious case of “dumbing down” here is the first part of *Percentages* where it is quite clear that the C1 version eliminated a step from the reasoning. However, the printing calculator technique does seem to work, and can therefore be extended to enable richer tasks such as *Taxi Times* which cannot easily be reduced to short answers to be presented on computer.

One unresolved issue is the case of *Triangle* where the C2 printing calculator variant appeared to be substantially easier than the number-only C1 version. This did not appear to be due to the availability of part marks, so it can only be surmised that the effect was either

“luck of the draw” or that the novelty of the printing calculator simply made an otherwise dry geometry task more engaging.

It is possible that the calculator adds a new, presentational step to the task: by design, rather than try to invisibly log the student's actions, it requires them to actively choose a printout and attach it to their answer. This may not be strictly equivalent to providing “space for working” where it is for the marker to extract evidence from the rough work.

The graph drawing tool

The *Lines* task used this in the most traditional way – to plot two equations on a graph and record their point of intersection. However, in this case the C1 version – a variant of multiple choice using drag-and-drop labels to identify the correct lines – performed very similarly, and was considerably easier to implement and mark.

The tool proved more useful in *Trip* (discussed in detail earlier) where the elimination of the distance/time graph drawing sub-task from the C1 version clearly removed an element of performance from the question. The tool also enabled the use of richer tasks such as *Taxi Times* and *Eggs* which relied on the tool to enable the input of “lines of best fit”.

In addition, the graph drawing tool could be put to use as a simple diagram drawing tool in tasks such as *Sofa*. While that particular task proved rather too difficult for the sample, many students did manage to produce (mostly incorrect) diagrams and there is anecdotal evidence (Figure 6.23) to suggest that inability to use the drawing tool was not preventing pupils from engaging with the problem.

Markscheme		Response	C	M
1.1	Any piece 110x90 appears			<input type="checkbox"/>
1.2	Any piece 110x65 appears			<input type="checkbox"/>
1.3	Whole sheet cut lengthwise into 110cm and 90cm pieces			<input type="checkbox"/>
1.4	Whole sheet cut widthwise into 90 and 130cm pieces			<input type="checkbox"/>
1.5	Fully correct (any response that gives the required pieces)			<input type="checkbox"/>

Figure 6.23: Pupils successfully engaged with the drawing tool - if not the task itself.

Marking Tools

The markers were familiar with online marking or data entry tools from their work with examination boards, and had little difficulty using this variant, especially after some refinements were made based on their feedback during the first round of marking.

The unusual aspect of this system is not the online entry of marks, but the presentation of responses to a computer based test (including richer responses such as the the graphs and printing calculator results) for manual marking (or confirmation of computer marks). The markers were able to mark responses based on these displays. In some cases, such as “line of best fit” questions the displays were enhanced with overlays to show the correct range of responses (see e.g. Figure 6.21).

Capturing marks for individual points on the mark-scheme, rather than just the aggregate score for each task, also proved invaluable as it allowed more detailed comparisons of performance, as shown in the previous section, even when the versions of tasks being compared had different structures or additional parts.

The computer-based marking was, in general, effective given the limited “proof of concept” ambitions for this project.

What was not tried here – but which would be a good subject for future study - was the “hybrid” model in which markers review the computer's marks and supplement them with any marks which can not be awarded automatically. Although most of the figures shown here include both sets of marks, the markers themselves did not see the computer-awarded scores.

Auto Marking

Unsurprisingly, the simple-answer C1 variants were marked reliably by the computer, with the only significant errors arising from the crude rules used for marking the textual “explain or describe” answers. The same can be said for those parts of the C2 questions with simple numerical answers, including crediting follow-through⁴¹ from previous answers (such as the correct co-ordinates for the point of intersection of two lines previously drawn on a graph). The main discrepancies were borderline cases, often representing a judgement which would normally have to be taken to moderation. For example, the answer to a currency calculation was \$569.1056911 and the mark scheme also accepted the nearest-cent value of \$596.11. Should \$596.1 also be accepted, given that currency amounts are usually written to two decimal places? Human markers thought yes – but the computer had been programmed to accept 569.10-569.11 with a minimum of 2 decimal places.

41 In marking terms, “follow through” means giving partial credit to a pupil who applies the correct method to incorrect results from an earlier step of the task.

6 - Design and evaluation of a prototype eAssessment system

In the case of the text answers, experimentation showed that even the simple system used here could potentially be refined to correctly mark all the responses encountered in the trial, but this would have then had to be properly validated using a larger corpus of sample answers including many more borderline cases. This was not a priority as more sophisticated ways of marking short textual answers have been well researched elsewhere (Leacock, Chodorow, 2003) and would probably be employed in a real examination system.

Auto-marking of simple graphs produced using the drawing tool proved successful: in the case of the *Lines* and *Trip* tasks the computer seemed somewhat more reliable than the humans at correctly identifying responses meeting the criteria, with several cases of markers failing to spot correct lines. Several borderline cases could have been resolved by refining both the human mark scheme and the marking algorithm. A number of computer mis-marks in *Eggs* were attributable to the arbitrary tolerances given to the computer to identify a valid line of best fit (ranges for the gradient and y-intercept) not exactly matching the guidance overlay given to markers (a shaded zone on the graph) in borderline cases.

Much of the auto-marking of the printing calculator tool was untested because, by GCSE conventions, most of the method marks were awarded by default on sight of a correct final answer. However, in those few cases where partial marks were significant the computer was able to apply simple rules to locate particular calculations or results in the working and give credit with a roughly comparable level of accuracy to human markers. The main source of discrepancy was in *Balls* which allowed a follow-through mark with which the current version of the software could not cope⁴².

More usefully, the trials showed that auto-marking could be successfully applied in tasks such as *Van Hire*, *Glass* and *Percentages* where working was required to validate a yes/no or true/false answer. With the first two, the computer and human marking compared well with most of the discrepancies being human errors or judgement calls. *Percentages*, as discussed above, showed a tendency for the markers to be over-generous in crediting mistakes not allowed for in the mark scheme.

Investigating discrepancies between human and computer marking not only identified technical shortfalls (usually repairable) in the computer marking algorithms, but a roughly comparable number of errors and inconsistencies in the human marking. Some of these were simple human error, but others were mistakes, ambiguities or omissions in the mark scheme which could affect both human and computer marking. Such issues were usually exposed where the marker made a common sense judgement which was at odds with the computer's strict interpretation of the rules.

⁴² The software can spot follow-through based on a value from a previous answer, but not currently from a value spotted elsewhere in the working.

6 - Design and evaluation of a prototype eAssessment system

An example of this was in *Van Hire* where the cost of using one van hire company could be interpolated from the table given in the question, whereas the mark scheme presumed that it would be calculated from the cost-per-mile worked out in the previous part. Some markers credited this valid alternative method and hence disagreed with the computer. However, adding this method to the computer's rules and re-marking reversed the situation, and exposed that the human markers had not been consistent in crediting the alternative method.

In a real examination, such inconsistencies would be picked up and corrected either during pre-test trials of the examination or during the normal process of marking and moderating the live examination. It is important to note that, with mark schemes of this complexity, such a checking process would clearly still be needed even with an entirely computer-marked test.

It was clear from the trials that even the level of sophistication of automatic marking used here could, as an absolute minimum, provide an invaluable cross-check on the accuracy of human marking. As a stand-alone solution for marking, the systems would need some further refinement and extensive testing against a large corpus of sample answers.

Where substantial pre-test trials are routinely constructed these last two points need not be an issue – since the trials will generate a corpus of sample answers and there will be an opportunity to identify problems, refine mark schemes, re-specify, re-implement and re-test the marking algorithms. However, for annual examinations that currently rely heavily on moderation during the marking and awarding process (i.e. GCSE) this could present a logistical problem, especially if the ambition was to provide instant results to candidates.

6.10: Feedback from pupils

No formal survey of pupils' attitudes was planned. This was partly because of the demands on pupils' time - they were already being asked to give up two lessons – and partly because the experience with both *World Class Tests* and *Progress in Maths* was that pupils were generally positive and uncritical about the idea of computer-based testing. However, the closing screen of the test (Figure 6.24) did invite pupils to type in their comments. About 60% of the pupils chose to enter a comment.

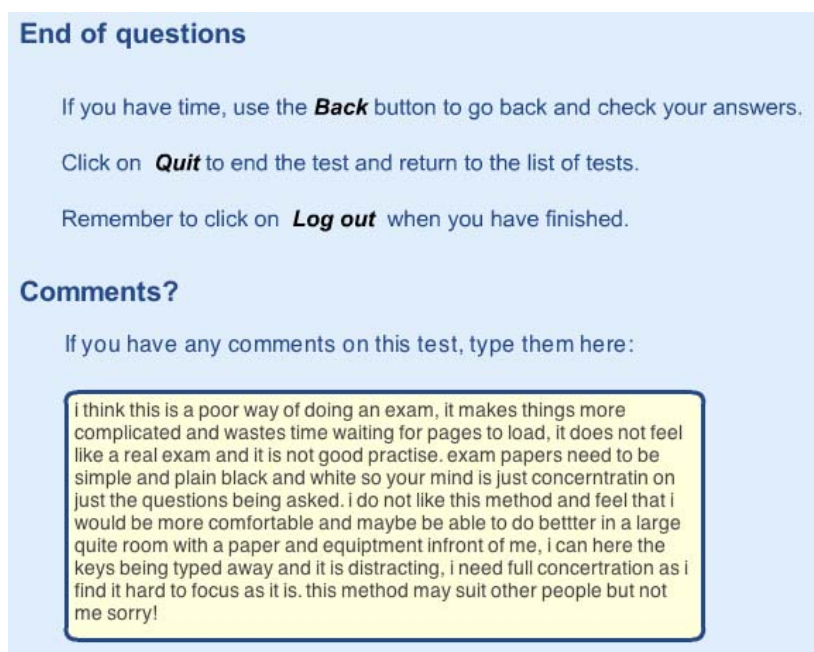


Figure 6.24: Feedback screen (including comment)

Table 6.13 summarises the results of a subjective assessment of the comments. Here, “general attitude” represents a judgement of the pupil's overall opinion, while “specific comments” summarises references to particular aspects or features of the system. This distinction arises when, for example, a pupil simply makes one very specific criticism (e.g. that the calculator needed a percentage key) without giving any evidence of their overall impression, or makes a general statement that they liked the test, without referring to details. The “comment quality” distinguishes vague statements such as “it was too hard” from more thoughtful responses which express a point of view (e.g. Figure 6.24) or describe specific issues.

6 - Design and evaluation of a prototype eAssessment system

It can be seen that although around 30% of the comments were generally positive, they were outweighed by negative comments. The strong bias towards specific negative comments is inevitable, because even pupils with a generally positive attitude made some criticisms and comments about specific features, while few singled particular features out for praise. Figure 6.25 shows the general attitude of pupils' comments broken down by their Key Stage 3 level: although this suggests a general trend towards more positive responses from levels 6-7 (it is to be expected that the difficulty of the mathematics would affect attitudes to the test) it does not indicate that disapproval is confined to lower-ability pupils.

N=163	General attitude	Specific comments	Comment quality	
Strongly negative	18%	13%	Frivolous or indecipherable	7%
Mostly negative	25%	39%	Vague, impressionistic	40%
Neutral or no evidence	27%	45%	Some supporting reasoning	39%
Mostly positive	17%	1%	Good – thoughtful or critical	14%
Strongly positive	12%	2%		

Table 6.13: Summary of pupil feedback

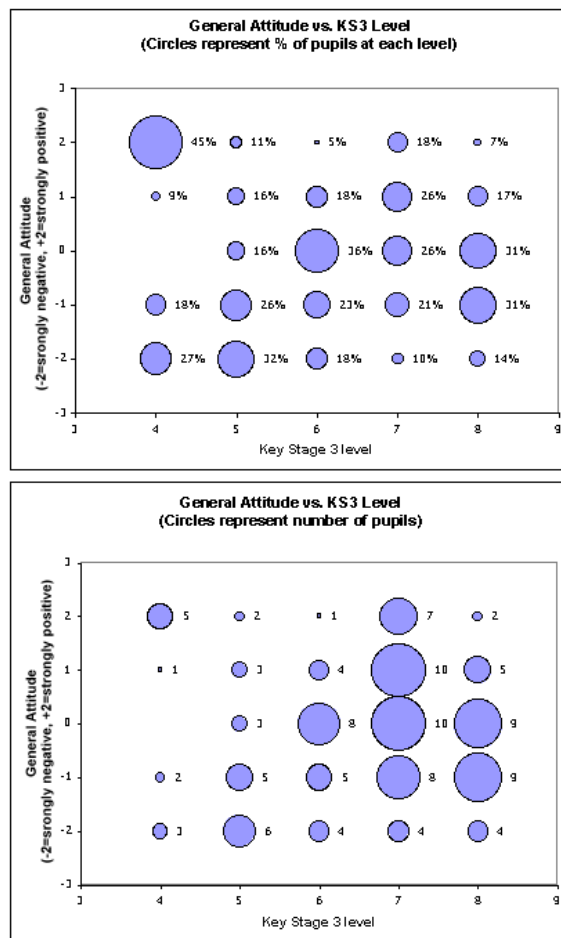


Figure 6.25: General attitude vs. Key Stage 3 level (by percentage and number)

Type of comment (comments may be classified under more than one type)	KS3 Maths Level						Total (out of 163)
	N/A	4	5	6	7	8	
Test was hard	10	2	7	4	6	3	32
Paper tests are better/easier	6	2	4	4	7	6	31
Disliked graph/drawing tool	12	0	0	4	5	8	29
Disliked calculator tool	4	2	1	3	4	8	22
Test was easy	5	0	1	3	8	4	21
Computer tests are better/easier	2	5	2	1	6	0	16
Technical problems	4	1	1	4	1	5	16
Computer distracted/reduced confidence	0	1	2	2	4	2	11
Liked calculator tool	0	1	1	0	1	2	5
Liked graph/drawing tool	1	0	1	0	1	0	3

Table 6.14: Frequent comments

Some of the recurring comment types are discussed below, with examples. The numbers of each type of comment, broken down by ability level (Key Stage 3 Mathematics level, where available) are summarised in Table 6.14.

“Paper tests are better/easier”: 19% of comments expressed or strongly implied a preference for paper tests over computer tests, or thought that paper tests were easier (sometimes this clearly referred to the paper test taken as part of this trial, others appeared to mean paper tests in general). While some of these were clearly expressing frustration over a specific software problem, some indicated more general issues:

To be honest i would probably rather do a paper test, this is because it is easier to show your calculations. It is said that you may use rough paper to do working out, but then you might as well be doing the test on paper anyway.

However it was a different way of doing a test which was more exciting than paper!

Pupil ID 490 (KS3 Level 8, Test 2A)

i found that this test was harder than the normall⁴³ test on paper because the ones on paper give you more help e.g with the drawing the line graphs and stuff and its not as difficult to draw and it is more easy to explin and do maths on the paper tests.

Pupil ID 609 (KS3 Level n/a, Test 2B)

If given the choice I'd rather use the paper version, beacause I found it easier to use and show your calculations.

Pupil ID 511 (KS3 Level 7, Test 2B)

⁴³ All pupil comments have been reproduced as written

6 - Design and evaluation of a prototype eAssessment system

As the test was computer based i found myself struggling with using the software which took my time up when doing the test. The calculator was difficult to use and you have to fap around with a mouse instead of a pen. I much prefer using paper and pen, as you can work things out quicker and simpler than using that calculator but it was a nice idea, and the variety colours made it interesting.

Pupil ID 491 (KS3 Level 7, Test 2B)

There is no way of putting notes on the actual diagrams, and so can be hard to think how to work the answer out. Also the calculator is harder to use than a real one.

ID 496 (KS3 Level 8, Test 2A)

“Computer tests are better/easier”: 10% expressed a preference for computer-based tests over paper. Recurring reasons given included neatness and clarity (either of the presentation of the questions or the entry of the pupil's answers), motivation or simply that it was “easier”. Reaction to the new tools was mixed – with some criticism even from those in favour of computer tests:

better than written tests as not as messy with writing everywhere

ID 180 (KS3 Level 5, Test 1A)

The test was very clear throughout. I personally felt it was easier then the paper test and would prefer to sit the exams using this technology. It was clearer then the paper test as it gave clearer, stepby step instructions allowing me to continue through the test more easliy.

ID 418 (KS3 Level 4, Test 2A)

i think this test is a good idea for future years becasue soke people cocentrate more when looking at a computer screen i find it easier to work on a computer and plus you don't have to worry about if your letters are clear enough.

ID 416 (KS3 Level 4, 2A)

I think the test is less boring than a real test and i found it easier to express my answers

ID 645 (KS3 Level 7, Test 1B)

It is a better system than doing a test on paper, but the fact that you still need rough paper is not good. I didnt like the way of taking printouts from the calculator, as when you put it in the box then if you already had an answer there it deleted the previous one and i wanted both sheets in it.

ID 546 (KS3 Level n/a, Test 2A)

i like this way of being assessed better than the old way of just doin a test on papoer. it makes you wants to do the test more because it is on the computer. however i do think that people will have to have a few practises on this kind of a test because it is a very different layout to what we are used to. overall i think this is a really good way to test people in maths!

ID 547 (KS3 Level n/a, Test 2B)

6 - Design and evaluation of a prototype eAssessment system

“Too hard” or “Too Easy”: 20% thought the test was “hard” compared with 13% who thought the test was “easy”.

Alot of the questions were hard to understand and possibly they could be improved by being wordered more simply.

ID 632 (KS3 Level 7, Test 1A)

the test is simple and the questions are explained well, and there ae not too many questions so youi don't get bored

ID 664 (KS3 Level 6, Test 1A)

some bits were really hard others were really easy!

ID 174 (KS3 Level 6, Test 1A)

i dont like it it was to hard!!!!!!

ID 205 (KS3 Level 4, Test 1B)

WAS HARD TO UNDERSTAND. AND IVE GOT ARTHURITUS IN ME FINGER.

ID 480 (KS3 Level n/a, Test 1A)

Predictably, the “easy” responses are skewed towards the level 7/8 pupils, while the complaints about the test being “hard” were more widely spread. There is an ambiguity about whether some of these comments are about the mathematical difficulty of the questions, problems with operating the test or a combination of both.

Computer distracted or reduced confidence : A small number of pupils (7%) specifically reported that they found the computer stressful, distracting or lacked confidence. For example:

also whats wrong with paper. and manual labour. i dont know what is trying to be proved by using computers, but mine started flashing purple and things and went fuzzy and put me off from answering questions. this WAS NOT HELPFULL you made me very stressed, although it did make me chuckle.

ID 157 (KS3 Level 6, Test 1B)

i couldn't concentrate as hard as i can when im working on paper and the calculator was kind of hard to use but apart from that it was quite easy to use. it was a bit harder to keep focused though i don't know probably because doing it on a computer was distracting

ID 354 (KS3 Level 8, Test 2A)

i think this is a poor way of doing an exam, it makes things more complicated and wastes time waiting for pages to load, it does not feel like a real exam and it is not good practise. exam papers need to be simple and plain black and white so your mind is just concerntratin on just the questions being asked. i do not like this method and feel that i would be more comfortable and maybe be able to do bettter in a large quite room with a paper and equiptment infront of me, i can here the keys being typed away and it is distracting, i need full concertration as i find it hard to focus as it is. this method may suit other people but not me sorry!

ID 156 (KS 3 Level 8, Test 1A)

The problems take too long to load. The coloured screen distracts you from the actually problem. It's hard to concentrate with the bright screen. It takes a lot longer to move the mouse and type the numbers into the calculator.

ID 353 (KS3 Level 7, Test 2B)

Negative reaction to graph and calculator tools: these reveal room for improvement in the design and ease-of-use of the tools, better instructions and some pre-test practice on their use.

It is a better system than doing a test on paper, but the fact that you still need rough paper is not good. I didn't like the way of taking printouts from the calculator, as when you put it in the box then if you already had an answer there it deleted the previous one and I wanted both sheets in it.

ID 546 (KS3 Level n/a, Test 2A)

I feel that some of the questions were extremely poorly written (for example q1⁴⁴), I also thought that it would save time if you didn't have that calculator as it took too long and wasn't that easy to control. Lastly I thought the graph questions were annoying as you couldn't plot the points to what you wanted and the lines were hard to draw with the mouse.

ID 543 (KS3 Level n/a, Test 2B)

Faults with the system: about 10% of comments referred to problems with the system, the most common being that questions took too long to download over the school internet connection, in severe cases causing errors.

Some problems may have been aggravated by poor/faulty equipment, such as worn or dirty mice or badly adjusted equipment. It was observed at one school, for instance, that some of the monitors were faulty or wrongly configured, making the grid lines on the graph questions almost invisible.

*My note paper didn't work :(
I found it too hard to read the writing
The squares on the last question were way too hard to see and hurt my eyes :(
And I didn't like it on the computer to be honest I like doing my work on paper.*

ID 481 (KS3 Level n/a, Test 1B)

the calculator was difficult to use. the mouse kept making the lines draw all over the screen- which took up lots of time. the calculator answers kept dropping on the answers I'd already written and didn't want to delete. I didn't understand how to work out some of the questions either as they weren't very clear. it was easier using paper on the first test, as you didn't have to keep changing from the computer to paper.

ID 500 (KS3 Level 8, Test 2A)

44 Q1 on this test was *Balls C2* – this could refer to a minor (but insignificant) typographical error, or simply show that there is no role for satire in mathematics exams (the “realistic context” was that a mathematics teacher wanted to know the answer).

The value of feedback

The feedback question was added as an afterthought, and was not a major focus of the research. Some of the responses suggest that this was a mistake, and a proper follow-up questionnaire would have been informative, although the added demand on volunteer teachers and pupils would have been an issue. Apart from the shortcomings of unprompted, unstructured comments as a way of comparing opinions, any survey on attitudes towards the use of Information Technology which can only be responded to by successfully using the same technology should be treated with caution: In this case, only pupils who successfully navigated to the end of the test would have the opportunity to respond, so some pupils who ran out of time, gave up, encountered software problems or simply disliked typing could be disproportionately excluded. So, had these responses indicated a glowingly positive attitude towards online testing, some scepticism would have been in order. Conversely, without any structured questions, it is possible that pupils who were frustrated by particular problems were more inclined to vent their spleen than those who had a neutral or mildly positive experience (the positive or neutral comments that were left tended to be shorter and less entertaining⁴⁵).

Some of the detailed feedback is invaluable for future refinements of the system: many of these anecdotes reveal real technical issues that could be addressed by design changes or improved support, and which might have been missed by a more quantitative, and possibly reductive, approach.

The unexpected result of this exercise was the number of strongly negative or critical comments: while many of these could be attributed to specific problems and failures of this system, there was also evidence of general scepticism towards computer-based testing of mathematics and anecdotal support for computers as being perceived as a distraction, rather than an aid, to GCSE mathematics. The contrast between this and the strongly positive attitude of the younger children towards the *Progress in Maths* tests (Chapter 4) - which were not without their problems and frustrations – is striking.

6.11: Practical and technical issues for schools

Online Delivery

Delivering software to schools in a form which they can easily install and use is an increasing challenge, especially when asking schools to voluntarily donate their time to a research project (a high-stakes awarding body could afford to take a somewhat more assertive approach).

⁴⁵ The shortest comment being “meh” - which the author believes is roughly synonymous with “whatever” or “am I bothered?”, probably counts as positive in the context of teenagers' attitudes towards maths.

6 - Design and evaluation of a prototype eAssessment system

Most schools have networked systems with fairly tight security that prevents pupils, and sometimes teachers, from simply inserting a disc and running or installing software. Installing software centrally on a network requires the co-operation of over-stretched (and often, not over-qualified) IT support staff. This makes a web browser based system particularly attractive, since it can be safely assumed that all school PCs have a web browser, and that most will also have the ubiquitous Flash Player “plugin” required by this system. All teachers need to do is to provide pupils with the web address and login details.

The other attractions to an entirely web-based system are logistics and security ⁴⁶: none of the test content needs to be physically delivered to schools ahead of time, nor can it be retained by schools after the test.

The disadvantage of this approach is that multiple brands and versions of web browsers (and the Flash player) are in circulation, with new versions or “critical updates” being released continuously. There are also security settings within the browser which can disrupt tests by generating warning messages whenever interactive content is accessed. This requires continuous testing and troubleshooting by the test developer, whereas a “bespoke”, stand-alone program can be constructed with fewer dependencies on other software.

There is also a lack of control over how browser-based tests will be presented: normally, browser content is displayed in a regular window with controls, such as next/back and a “close window” button, which could disrupt the test if used. Efforts to hide or disable these, or to make the test window fill the screen, are particularly vulnerable to variability between browsers and security settings (this was noted during the observations of the browser-based *Progress in Maths* tests).

In an attempt to cover most eventualities, trial schools were offered a choice of three delivery mechanisms:

- A. **Browser based:** as described above – if users had a suitable browser and plug-ins they could run the test by simply visiting the web site and logging in.
- B. **Online player:** a minimal application “shell” that could be downloaded and used to run the online tests in a well-defined environment without depending on a web browser and plug-ins. All of the content was downloaded from the web as needed, and not stored on the pupils' computers.
- C. **Full install:** this allows the shell, tools and all the required tasks to be installed on the pupil's computer or a local disc. The internet connection is still required for authentication, control and the return of pupil responses.

⁴⁶ No claims are made for the security of the prototype used here: any system used for delivering real high-stakes tests would need detailed analysis by security experts, particularly if it was required to keep the test content secure after the test had been taken.

6 - Design and evaluation of a prototype eAssessment system

Trial schools were offered all three methods (via CD or as a download) with, initially, a strong recommendation for option B, which was seen as the best compromise between a controlled environment and online delivery.

During the first round of trials, however, reports emerged from two of the four participating schools of frustrating waits of several minutes between questions or software time-outs. This was unanticipated since the actual volume of data downloaded for each question was roughly equivalent to (for example) opening the “BBC News” home page and, when tested on a typical home broadband connection, loaded within a few seconds. For the second round, the recommendation was changed to method C, but there were still problems with speed.

While it is possible that these problems were caused or exacerbated by the central server ⁴⁷, or some design flaw in the server software, only some schools were affected and so it is suspected that the bottleneck was caused by schools' own internet connections. All schools had been told that they would need a fast internet connection to take part, and to consult their technicians. After the event, the technician at the school with the worst problems characterised that school's internet connection as “very slow”. Another school (which didn't eventually take the tests) reported that:

“Our school has one 1 meg (Mb/s) connection for 1500 people (As you know, schools have no choice of ISP)”

This (at the time) was the type of domestic internet connection being offered for £20/month. Another possible culprit is the widespread use of “proxy” servers for content filtering, often run by the local education authority or their appointed internet provider: in theory, these should speed up access by keeping caches of frequently used pages, but in practice, if they are not sufficiently powerful to match demand, they will act as bottlenecks.

Another issue was highlighted during one school visit to observe the tests: the internet connection failed just before the session and there appeared to be no mechanism for having the fault investigated or obtaining an estimated time in which it would be fixed.

One school deserves a heroic mention:

“One of the e-desk rooms was invaded by plumbers 20mins before we were due to start and we had to relocate. We had a power cut 5mins before we started.”

One assumes that had the test been a real GCSE, the plumbers would have been the ones relocated, but this does illustrate that contingency planning (and back-up generators) would be essential for any high-stakes computer-based assessment.

⁴⁷ This was running on dedicated hardware with the type of fast internet connection enjoyed by a large university, and was seldom dealing with more than 20 or so students at a time.

At the time of the trials, therefore, although online delivery of tests offered significant logistic benefits, the provision of broadband internet at many schools was not adequate for this purpose, and that internet and ICT provision was not always regarded as a “mission critical” service. However, this is a rapidly developing field, and schools are increasingly using the internet as a central part of their teaching so it is possible that this situation has changed since the trials were conducted. In particular, the maximum bandwidth of domestic broadband is continually improving, with more options offering guaranteed bandwidth for non-domestic uses.

Other practical issues

There were few reported problems other than the slow internet connections and occasional vaguely described “crashes” which were hard to attribute to any specific cause.

Other potential problems noticed during visits to participating schools include:

- The potential for copying, when the computer room layout allowed pupils to easily see each other's screens.
- Adjustment of displays: the specific problem observed was that some screens did not display the blue grid lines on the graph questions⁴⁸. Other potential problems could be if displays were configured to stretch or squash images (likely with newer “widescreen” displays).
- Mice: Some pupils complained of difficulty drawing accurately with the mouse: while this may be a software design issue, faulty mice, unsuitable surfaces or threadbare mouse mats could also contribute. Aside from physical problems, many aspects of mouse behaviour (such as the sensitivity, or the functions of the buttons) can be adjusted by the user.
- Working space: even with tools for showing working, rough paper can be invaluable when doing mathematics. Pupil need space for a keyboard, mouse **and** somewhere to write.

The other, major, issue is that even those schools with good IT provision would not have had sufficient computers to enable an entire year group to take online examinations at the same time. Nor would they have had sufficient support staff to provide the level of technical support that pupils taking life-changing high-stakes tests would expect.

⁴⁸ Although all modern PCs are capable of resolving “millions of colours” they can easily be set to modes with fewer colours. Alternatively, the screens in use may have been faulty or simply badly adjusted: the grid lines are intended to be faint, but visible.

6.12: Conclusions

The work described in this chapter set out to address research question B: (Section 1.2) - “What are the effects of transforming an existing paper-based test to computer?” and D: “What are the implications for the technical and pedagogical processes of computer-based assessment design” in the context of a design research project. Whereas other studies (such as *NAEP, 2005*) have done this for short answer tests from the USA, the tasks chosen here were mostly “clones” of typical GCSE mathematics tasks, along with a few examples taken from *Balanced Assessment* and *World Class Tests* which attempted to incorporate slightly longer chains of reasoning. These tasks included elements such as space for showing working and drawing simple graphs and diagrams which were not found on the USA-style tasks. Suitable tools had to be designed to enable these task types.

Although no attempt was made to adequately sample a complete syllabus, the intention was to try a diversity of task types rather than to study a single task in detail.

As well as the paper vs. computer comparison, the study compared two models of “task transformation” - one (the “C2” model) attempted to use the new tools to reproduce the paper version as closely as possible, while the other (“C1”) re-wrote the task to a short-answer form which could be delivered and marked using most existing computer-based testing systems.

The capture and marking of working presents a technical challenge for computer-based testing. If – as in many GCSE tasks – it is only to be used to provide partial credit when the candidate makes an arithmetical error in a calculation then the results here suggest that only a few percent of candidates benefit directly from these partial marks. However, in other cases, it enables the use of questions with longer chains of reasoning (such as the first part of *Percentages*) for which the more structured, short-answer version is unsurprisingly easier. There are also cases (*Triangle, Currency*) where pupils appear to benefit indirectly from being asked to show working, perhaps because it increases concentration or engagement with the question. This should be balanced by the anecdotal evidence from the feedback that pupils found having to use the “printing calculator” tool frustrating.

The value of the graph drawing tool also seems to depend on the question – there is little difference between the short answer/multiple choice version of *Lines* and the version with the drawing tool. However, it was not clear how it could have been replaced for drawing lines of best fit in questions such as *Eggs* while its use in *Trip* revealed a major misconception about gradient which the short answer failed to expose.

One practical finding was the difficulty of recruiting and retaining sufficient volunteer schools to perform the type of cross-over study envisaged here without substantially larger

6 - Design and evaluation of a prototype eAssessment system

resources and the risk of exploiting the good will, generosity and patience of teachers⁴⁹. Consequently, few quantitative conclusions can be drawn about any systematic effect of computer versus paper tests. The data comparing “simple (C1)” and “rich (C2)” tasks, randomly assigned within the same classes (easily achieved with computer delivery), is more substantial.

Although the experiment proved too coarse to reveal any overarching effect on difficulty caused by the modes of presentation, it was valuable as a design research exercise. In particular, it shows how the capabilities of an eAssessment system can influence the type of questions that can be asked, and hence the potential balance of the test. While many GCSE tasks can be adapted to short answer versions without the need for “rich response” tools, others would have to be changed considerably. Furthermore, adopting a system for GCSE without rich response facilities could lock the assessed curriculum to the current state of GCSE, with the shortcomings discussed in the previous chapter, making it difficult to broaden the range of task types in the future.

In terms of automatic marking, the priority here was to capture responses in a consistent, computer readable form, on the assumption that once data had been captured, marking algorithms of arbitrary complexity could later be developed and tested against a corpus of scored responses. The marking algorithms actually used in the trials were mainly “proofs of concept”. However, even the techniques used were comparable in accuracy to the human markers, and would, at the least, have value as a way of checking human marking. Even if automatic marking of these tasks could not be perfected, the “manual” marking system adopted proved efficient and, while entirely online and paperless, could still present pupils' working and drawings to markers.

⁴⁹ It is worth noting that a similar problem afflicted the Progress in Mathematics equating study analysed in Chapter 4 - where a respectable number of individual participants was subverted by a small number of schools.

7: Conclusions

Johnny had written “What it felt like to be different sorts of peasants” on it and printed them out on the printer, although he had to rewrite them in his handwriting because although the school taught Keyboard Skills and New Technology you got into trouble if you used Keyboard Skills and New Technology actually to do anything. Funnily enough, it wasn't much good for maths... they wouldn't let you get away with “what it feels like to be x^2 ” ...

Terry Pratchett – 'Only You can Save Mankind'

7.1: Introduction

In this section, we start with a review of the original research questions, drawing together the conclusions of the various strands of the study, and suggest some areas for continuing research.

Finally, we discuss a general question which permeates these results: whether the computer is, for pupils, a “natural medium for doing mathematics” and whether current assessment reflects and values the role of computer technology in modern mathematics.

7.2: The research questions revisited

Research question A

How can eAssessment contribute to the assessment of problem solving skills in mathematics?

The work on the *World Class Tests* (Chapter 3) illustrated the potential role that the computer could play as a presentation medium for tasks set in richer, realistic contexts. The potential advantages of computer presentation include:

- The use of animation or interactive graphics to communicate concepts and information which would be hard to communicate, in simple language, on paper
- The provision of a substantial data set, for students to explore with searching or graphing tools
- Simulated experiments, games and other “microworlds” - allowing question types that would be impossible on paper

Experience during the development of these tests, supported by other studies, shows that the computer could allow pupils to engage with such tasks, even when the subject material was completely outside the taught curriculum. Examples of this include the task on Archimedes' principle for 9 year-olds (Figure 3.2, p36) and multi-variable problems such as *Oxygen* (Figure 3.8, p41) which were further investigated in another study (Ridgway et al., 2006).

The problem solving content of *Progress in Maths* (Chapter 4) appears limited in comparison with *World Class Tests*. One factor is that open, extended problems tend to require quite substantial “prompts” to introduce the context and present the available data, which is a problem when designing tasks to be taken in a test environment by young children with a wide divergence of reading and language comprehension skills. This is a limitation of the formal test format: it is to be hoped that children as young as 6 or 7 will encounter authentic problem solving in the classroom, or as part of face-to-face teacher assessment, where there is a teacher on hand to help them understand and engage with the task. The only caveat here is that some routine tasks in PIM appeared to have been promoted as problem solving in the teachers' guide, when they could easily be answered correctly by ignoring the context and performing the obvious exercise by counting the objects or doing the sum. Most of these issues, however, were common to the paper and computer versions of *Progress in Maths*.

Although the GCSE-level study concentrated on more conventional questions, a few tasks adapted from paper tests with stronger problem-solving elements were included. These proved somewhat too challenging for the pupils involved, both on paper and computer, although it does appear that those students who did engage with the questions were also able

to use the calculator and drawing tools to present their responses. It is possible that these tasks could have been made more accessible using *World Class Tests* techniques to present the context in a more engaging way (such as explanatory or context-setting animations) but that would have invalidated the trials' initial objective of comparing performance between computer and paper versions. Including such tasks could help GCSE to assess more of its intended objectives and to meet new demands.

Research question B

What are the effects of transforming an existing paper-based test to computer?

Some previous studies (Sandene et al., 2005; Russel, 1999) have suggested an effect that gives rise to measurably lowered scores on computer-based mathematics tests, particularly when non-multiple choice questions are used. However, these studies concentrate mainly on aggregate performance across a complete test composed of short-answer questions.

The two studies described here – the *Progress in Maths* analysis and the GCSE-level eAssessment trial – failed to confirm the presence of a systematic effect, but did identify individual tasks whose difficulty seemed to be affected by the change of medium. This usually, but not always, meant that the computer version was harder. This would support the hypothesis (suggested in the NAEP report) that the changes in difficulty were caused by design changes to specific tasks necessitated by the translation process, rather than by some systematic computer effect. Several examples of the type of changes that might be involved, and their potential effect on the nature of the task were explored in section 5.4 (p125) and chapter 6.

That such changes might have an effect is hardly surprising: mathematics tasks can be highly sensitive to minor changes in presentation, as has been noted in studies of paper based assessments. For example:

*The operationalization, i.e., actual form taken by the question, chosen for checking a given skill is sensitive to such a degree that the change of only one word (even an article) can produce strong differences in the results. So, what can be said when the question is “dressed” differently?
What does To Assess mean? (Bodin, 1993)*

If there is a generic “computer effect” then it could be attributable to stress, distraction and additional cognitive load introduced by the computer, especially if pupils are required to use unfamiliar software or if they encounter technical glitches. The informal feedback from the GCSE-level trials revealed a surprising level of scepticism about computer-based testing, with stress and distraction a recurring theme (see section 6.10). While some of this was

7 - Conclusions

clearly directed at specific problems with the prototype test software used, other comments were more general and many users seemed much happier working on paper.

In the case of *Progress in Maths*, the younger children were, to the extent to which they could express themselves, happy with the idea of using a computer. However, apart from a few cases where specific design issues were spotted, any under-performance on the computer tests seemed to arise from an increased tendency to make random, silly mistakes. These tests, though, were something of a special case in that the paper tests were led by a teacher reading out the questions. Interviews with teachers who had delivered the paper tests revealed that they would often take a pro-active approach to ensure that their pupils listened carefully to each question before starting work on it. It is unsurprising, therefore, that a self-paced computer test based on recorded, spoken prompts would result in poorer concentration and failure to properly listen to questions.

Generally, though, the impression is that, where a task can be translated to computer without obvious changes to the required activity, then it will “test the same thing” as its paper counterpart, if not at exactly the same level of difficulty. However, it is unlikely that an entire existing test can be translated as-is without requiring some tasks to be redesigned, and these will require careful trialling. The prudent approach seems to be to pilot, validate and calibrate computer tests separately from paper tests, not to assume that they can inherit the data from existing paper tests.

Research question C

How might eAssessment be used to improve the range and balance of the assessed curriculum (and hence, indirectly, the taught curriculum)?

This is a crucial question for this thesis. To produce a computer-based test “as good as” existing assessments, such as GCSE, presents many challenges, both in terms of software design and providing adequate school infrastructure, but is undoubtedly soluble. The work here even suggests that some of the trickier-to-implement features of GCSE, such as “method marks”, are not well exploited by current tests and could probably be dropped without greatly impoverishing the syllabus. What, though, are the implications of this approach for the mathematics assessment reform movements discussed in Chapter 2?

A recurring theme of mathematics assessment reform is an increased emphasis on longer tasks involving problem solving or modelling skills and set in realistic contexts. The *World Class Tests* (Chapter 3) focussed on this aspect and showed how the new medium can aid the delivery of such tasks in a formal test environment. However, that project only began to tackle the issue of collecting and scoring the responses.

7 - Conclusions

Scoring problem solving tasks often requires markers to be able to distinguish between a range of levels of performances, not to simply mark the answers as right or wrong. Even where such a task does have a short and well defined correct answer it would rarely be acceptable to spend 10-20 minutes on a task for which the only feedback was “right” or “wrong”. The answer may be less important than the work which supports and justifies it and a completely correct answer might only be expected at the highest level of performance, with considerable credit available for partial responses.

The most generally applicable method of implementing such tasks is to require extended constructed response answers. So while the ability to collect and score “method” may be dispensable for current GCSE tasks, it would be an essential feature for any test which sought to be better aligned with the intended curriculum. The input tools discussed in Chapter 6 enabled some richer questions from other projects to be included in the GCSE trials, although these proved challenging to the intermediate-level pupils involved.

Without such a capability, it becomes necessary to split the task up into sub-tasks, so that each stage produces an answer which can be captured and marked. That is a viable solution for some tasks, but in other cases it can undermine the problem solving aspects of the task by leading the pupil step-by-step through the approved solution path and reducing the *unsupported reasoning length* of the task. Where this technique was used in *World Class Tests* it was for pedagogical reasons, to tune the overall difficulty of the task, not because the mark scheme demanded it.

While *World Class Tests* was free to invent its own curriculum, the rest of the work described here focussed on existing tests of the mainstream mathematics curriculum. Here, any change is likely to be evolutionary, rather than revolutionary. One of the common themes that emerged from this is that the mathematics curriculum tends to be at odds with the realistic use of computers in assessment. Some possible reasons for this are discussed in section 7.4. When the computer does not really play any role in the pupils' mathematical work, the art of computer-based assessment can become preoccupied with reproducing traditional tests on screen – something which is true of both the *Progress in Maths* tests and the GCSE-level trials described here. The main design consideration becomes how to capture mathematical responses without changing the nature of the task or requiring the pupil to deal with a complex user interface (see section 5.4, p125). While such design work is necessary to ensure that tests are not narrowed further by computerisation, it does not itself expand the assessed curriculum.

Ideally, then, the assessed curriculum needs to shift so as to reflect the central role of computers in modern mathematics, to allow and encourage pupils to use real-world mathematical software. If we could assume that all candidates had basic fluency in using a

7 - Conclusions

suite of such tools, and these could be made available during the test without distracting pupils from the tasks, then it would greatly simplify our problem of capturing mathematical reasoning in a computer readable form, and enable richer tests which focussed mathematical reasoning.

The challenge to this is that “real” mathematical software can often automate and trivialise tasks such as arithmetic and algebraic manipulation which are often central to current assessments. If these facilities are disabled, the tool may seem pointless, although some of its conventions and notations might still be used to collect responses.

Projects such as MathAssess (Fletcher, 2009) have tackled many of the technical challenges of presenting mathematical notation, capturing and scoring responses containing mathematical expressions. However, the emphasis appears to be on allowing “mathematical expressions” to be used as a format for a short answer, rather than on capturing and crediting reasoning. Tellingly, while MathAssess incorporates a computer algebra system to facilitate the scoring of responses, the report notes that candidates' input must be “limited with respect to what it can do when passed to (the computer algebra system)” to avoid using the system's functionality to “carry out algebraic manipulations which it is intended that the candidate should do” (Fletcher, 2009, p. 5).

It is worth reiterating that the “assessed curriculum” is not the same as the “intended curriculum” or the “taught curriculum”. At GCSE, for example, the intended, statutory, curriculum is defined by the National Curriculum for Key Stage 4, while the “assessed curriculum” is defined – officially – by the specifications published by the Awarding bodies. More realistically, the *de facto* assessed curriculum is defined by the past and specimen papers released by the awarding bodies, since only here will you find actual examples of the types of tasks which need to be mastered to pass the examination. In many classrooms, the “taught curriculum” will be closer to the “assessed” than the “intended” curriculum (see Section 2, p12).

Typically, the assessed curriculum is narrower than the intended curriculum: the National Curriculum certainly recognises more potential applications for IT and has a greater emphasis on “problem solving” and “process skills” than is reflected by the GCSE examination. Hence, there is considerable scope for innovative online tests to expand the assessed curriculum to better match the intended curriculum. Computer-based assessment can, potentially, regulate which tools are available at various stages throughout the test with greater finesse than can be done with (say) calculators during a paper test. Hence, making powerful, realistic tools available for rich questions does not preclude the assessment of basic skills during the same test session.

Research question D

What do the above issues imply for the technical and pedagogical processes of computer-based assessment design?

The findings in respect of question B, the effect of translating a task to computer, could be summarised as *details matter*. This will not surprise any designer of assessments who has been involved in the trial and refinement of tasks, but it might not be so obvious to an IT specialist charged with implementing the work of task designers. Some common types of changes between paper and screen, with potential for significantly changing the nature of the task, were noted during the analysis of *Progress in Maths*:

- The “prompts” for computer versions often became longer than the paper equivalents, because they required extra instructions on how to operate the computer
- Prompts could be subverted by changes on the computer screen: for example, one question asked pupils to “fill in the missing dots”, yet once the pupil had added some dots, there were no “missing” dots. It is hard to imagine this causing confusion on paper, but it was observed to cause problems on the computer
- “Improved” graphics could introduce new distractions. In *Progress in Maths* most tasks on paper were already illustrated, but the introductory instruction screens had new graphics added, and these were observed to cause serious distractions when pupils thought they saw mathematical problems in them
- Splitting questions over two screens – this requires careful thought as to which information needs to be presented on both screens. On the other hand, although the NAEP and Russel studies recognised multiple-screen questions as a problem, the *Progress in Maths* tests suggested that the second part of a two-part question could sometimes become easier when presented on a second screen
- Changes in the layout of graphical multiple-choice items, so that the eye was drawn to a different solution
- Revealing multiple choice options one at a time, so that pupils could see, and select, a distractor as their final answer without ever seeing the correct solution (or *vice-versa*)

These represent precisely the sort of changes which might have been made by a programmer when implementing a paper based task, if the original designer had not specified otherwise. Hence, it is essential that designers of tasks for computer-based assessment gain experience in designing for that medium, so that they can envisage, and specify in detail, the required presentations and interactions.

7 - Conclusions

The other essential element is that new tasks should be subject to small, closely observed trials in which the objective is to gather qualitative data on how pupils interact with the tasks, and to spot cases in which the task is not assessing the intended mathematics. Useful techniques include:

- Pupils working in pairs, to encourage them to externalise their thinking
- Observers intervening in pupils' work, asking probing questions or helping them operate software, to ascertain where the problem lies
- If possible, the ability to quickly modify and re-trial the tasks as soon as a design issue is identified

Clearly, though, these actions are incompatible with collecting impartial, quantitative data from a statistically significant sample. Hence it is important that these initial cycles of trial and refinement take place before a move to larger-scale quantitative trials.

While the same principle could be applied to paper tests, there are two factors which make it more salient to computer-based tests:

Firstly, when analysing the results of the GCSE-level trial, the pupils' written responses often had revealing comments and annotations, even where formal working had not been given. Regardless of whether this affected the pupils' ultimate performance on the task, such information is invaluable when trying to understand how pupils are engaging with the question. Even with tools such as the printing calculator, such clues were largely absent from computer responses, and could only have been obtained by "interactive observation". Issues will be even harder to spot on computer-only tasks where there is no paper version to compare performances against.

Secondly, whereas a paper task can easily be tried out by the author, in rough form, a computer task must usually be programmed before any trial can take place. In many projects, this would mean that the task had already been specified in detail and handed off to the software developers. The inevitable time pressures then encourage leaping directly to a large psychometric trial. By analogy, it is as if a paper task could never be shown to a pupil until it had reached the stage of a typeset printer's proof.

Thus, it is essential that computer-based test development allows sufficient time for informal trials and subsequent refinements, and recognises that this might take longer than for conventional trials.

7.3: Questions for future research

Further data collection

More trial data would be required to fully investigate whether there was a systematic difference in difficulty between supposedly equivalent paper and computer versions of the same test. The practical experiences in this work show that this would need to be on a larger scale, covering more schools and with tighter controls on ability levels to ensure a valid cross-over sample. This would be helped by random allocation of tests on a pupil-by-pupil basis, although that would complicate the administration.

However, since we have seen that most computer versions of paper tests will include new or modified tasks, and that the performance of these can change quite significantly, looking for a smaller, systematic test-wide effect might not be a high priority.

Pupils' mathematical IT skills

During the work on this, and other educational IT projects, some anecdotal evidence suggested that pupils' experience in using computers as a mathematical tool is limited, and that some mathematics teachers assume that skills such as the use of spreadsheets are the responsibility of ICT teachers. Further research in this area could be informative, as it pertains to the role of the computer as a “medium for doing mathematics” discussed above.

Some of the anecdotal evidence arose during the development of professional development materials for the *Bowland Maths* initiative (Bowland, 2008), specifically the reaction from teachers and advisors to the inclusion of a simple spreadsheet modelling task. In this material, three roles were suggested for computer use in mathematics classroom, which could be the starting point of a model for such a study:

1. A “thinking tool” for representing and analysing problems (this corresponds most closely to our “medium for doing mathematics” argument presented later in this chapter, a key property being the development of transferrable mathematical IT skills)
2. A “microworld” offering a rich domain to explore (several of the *World Class Tests* tasks typify this)
3. A “didactic tool” that explains and gives practice (*Progress in Maths* and most of the GCSE-level questions would fall under this heading, as would most whiteboard presentations and non-interactive animations used by teachers).

Note that this does not refer to the type of software, but to the **mode of use** in the lesson. Most mathematical tools (graphing, geometry, algebra, spreadsheets) can potentially be used in all three roles.

Assessment reform – in any medium

In England, major changes to GCSEs are in progress⁵⁰, including:

- Reducing the current 3-tier system (foundation, intermediate, higher levels) to a two-tier system
- Introducing more emphasis on “functional skills”
- Offering a double-subject mathematics qualification, with separate “methods” and “applications” components.

At the time of writing, these changes are now being piloted (see e.g. Murphy & Noyes, 2008) with the revised papers just becoming available. These developments will inevitably have implications for the issues discussed here, although the need for developing computer based versions of these new tests seems less urgent than in the earlier statements quoted in our introduction.

Formative vs. summative assessment

This work has concentrated on traditional, summative tests, which typically produce a score or grade for each pupil. There is increasing interest in *formative assessment* which provides more comprehensive, specific feedback to help pupils develop their skills and teachers to refine their teaching. A seminal piece of work in this field, *Inside the Black Box* (Black & Wiliam, 1998) showed that classroom assessment leading to qualitative feedback could substantially enhance student learning. However, the same work found that these gains were easily lost if the feedback was accompanied by a score or grade. The authors have since been critical of the way the name *formative assessment* “has been applied to regimes of frequent summative testing, which the original evidence for formative assessment does not support” (Black, 2008) .

Computer-based testing systems often make a selling point of their capabilities for generating and reporting summative statistics and standardised scores. To review what extent existing testing systems support the analysis of pupils' actual responses could be a relevant subject for study.

Use of alternative platforms

The time will surely come when all pupils have access to a computing device throughout the school day. What is less clear is whether this will be a traditional desktop or laptop computer, a mobile device, an advanced graphing calculator (possibly with algebraic and geometric capabilities) or a “smart pen” that can store and record writing.

50 http://www.qcda.gov.uk/24956.aspx#Background_to_the_pilot

7 - Conclusions

There are many ongoing research projects into the use of particular technologies in learning (a wide selection can usually be found at <http://www.lsri.nottingham.ac.uk/>). Such studies reveal a diverse range of, often, quite specific applications: making a video blog on a mobile phone camera, using SMS to feed back during a lecture, using “voting systems” for whole class teaching. Each offers a potentially valuable contribution to a balanced education, and some certainly merit investigation for their possible applications in assessment, particularly formative assessment.

However, unless there is a major abandonment of the current assessment model, summative assessment will need a platform that can deliver the whole of the school curriculum. It was seen in Section 2.2 (p12) how assessment shapes what is taught in the classroom, so the adoption of diverse technology in the classroom could be hampered if the same diversity is not present in assessment.

The attraction of the “traditional” general purpose computer is that it can potentially fill the roles of (or work alongside) all of these devices, and more. A very recent development is the commercial success of “netbooks” - these are very similar to (if not indistinguishable from) from regular laptops but have been designed with low price, compactness and network connectivity, rather than processing power, storage capacity or large high-resolution screens as a priority. These are an attractive proposition for schools looking to provide one-to-one computer access for their pupils.

Smart pens and “tablet computing” have been around for some years without achieving widespread adoption. Smart pens work exactly like normal pens and write on paper, but record everything written for later upload to a computer (and often offer handwriting recognition). One possibility is that they could be used alongside a computer-based test to collect working, without the extra expense of collecting and returning the paper. A tablet computer allows the user to write directly on the screen with a stylus, and could solve the problem of entering mathematical notation or free-form diagrams – however, full-size tablet computers have remained expensive, and the difficulty of writing with a stylus on small mobile devices might discourage their adoption. At the time of writing, touch-sensitive screens are becoming common in mobile devices, but these are finger-operated and tend to rely on on-screen buttons or keyboards for input, rather than allowing writing or drawing.

The development of these technologies is rapid, and their uptake will obviously have a bearing on the viability of computer-based assessment, especially using tasks which employ the computer as a tool.

7.4: The computer as a medium for “doing mathematics”

A new medium

An interactive computer program can be a very different medium to the printed page.

This seems obvious, but the attitude of publishers sometimes suggests that this change of medium is akin to a different weight of paper or the introduction of two-colour printing. The initial brief for *World Class Tests* called for an all-computer test, and potential software developers were interviewed and appointed quite independently of the test designers. The publishers of the *Progress in Maths* tests presumed that the new digital tests could be equated to the existing paper test via a much smaller study than the original calibration exercise⁵¹. All the studies cited in this thesis (including, to an extent, the new work described here) take as their “null hypothesis” that the expected performance on computer-based test is identical to the paper test it replaces.

Are these reasonable expectations, or is the real surprise that pupil performance on two such different media is often so similar? Could these expectations arise because, while many people regularly use computers for presenting and creating text, fewer have experience of using them as a mathematical tool?

Computers and writing

For most people who use a computer for writing, the *target* medium is still paper (or an electronic facsimile thereof). Most of the conventions for language and presentation of computer-written documents have been directly inherited from conventional publishing and writing, while software developers have devoted much effort to the faithful reproduction of the traditional printed page. Here, for example, we have a thesis discussing interactive computer software, yet it is still arranged in chapters, with page numbers, tables of contents, footnotes and a bibliography⁵². The same output could have been produced – albeit via a longer and more expensive process – 50 years ago.

So, for the writer, the computer acts as an improved tool for writing. It may enable new ways of working (authors might, for example, decide to apply typography as they write, cut and paste passages that they would previously have rewritten or just stop worrying about spelling and grammar) but since it is a tool designed to simulate traditional media it does not impose any changes on the fundamental nature of the task. It is not, for example, necessary to learn new rules for spelling, grammar and punctuation to enter text. While the computer might

51 Both parties were quite willing to change these positions in the light of evidence or reasoned argument, but it is interesting to note their initial assumptions.

52 If the author prevails, it may have colour illustrations and may not be double-lined spaced. There could even be a CD-ROM in the back. However, submitting it as online hypertext or a virtual gallery in *Second Life* might be seen as a “brave decision”.

improve the author's spelling, or point out the occasional grammatical error, it does not take over the task of composing the text. Conversely, a word processor with grammar and spelling checking removed (as might be desirable in an English test) is still a useful writing tool which works largely as expected.

There are now widely adopted, if unwritten, common conventions for how word processing applications work. The more advanced features and techniques may vary between different products, but the techniques for text entry, editing, cut/copy/paste and simple formatting are usually familiar.

New media such as multimedia and the World Wide Web – particularly the type of user-written and collaborative content now emerging (e.g. Crook, 2008) – represent a more radical change in medium. It may well be that, in the near future, the ability to create a hypertext-rich wiki article will become a prerequisite for employment. For the present, though, the ability to produce traditionally structured documents is the ubiquitous skill amongst computer users.

Hence, the computer has become a natural medium for writing, but one which largely replicates more traditional media and augments, rather than transforms, the act of writing.

Computers and mathematics

The relationship between the computer and mathematics is paradoxical. On the one hand, a computer is a fundamentally mathematical entity, conceived by mathematicians, whose design, study and programming form a branch of mathematics. However, particularly in the case of the personal computer, the majority of people who use computers are using them for writing, communication, graphics or data handling. They do not consider themselves to be “doing mathematics” - and would probably panic if anybody suggested that they were.

The only overtly mathematical tools which approach the ubiquity of the word processor are the on-screen calculator and the spreadsheet⁵³. These are versatile tools, but neither has the same broad application to mathematics as a word processor does to writing.

Most word processors do have facilities for entering mathematical notation (such as *Equation Editor* in Microsoft Office), but these are designed primarily as a way of statically representing already known mathematical expressions and lack any ability to manipulate or evaluate them.

To find a class of mathematical software with such a broad domain, it is necessary to look to generic programming languages or mathematical analysis tools such as *Derive* or

53 The author has seen enough children and adults using the former to calculate values for entry in the latter to question whether the spreadsheet's mathematical capabilities are widely appreciated.

7 - Conclusions

Mathematica which do offer their experienced users a versatile medium for doing mathematics. Alternatively, there are powerful tools for narrower domains, such as graph plotting, statistics, dynamic geometry or specific modelling techniques. However, such tools tend to be designed by and for mathematicians, engineers and scientists. Where these are used in schools, the typical application is more likely to be as a delivery system for expert-written interactive demonstrations⁵⁴ rather than an open-ended tool for pupil use. Becoming fluent in such a system requires a considerable time investment, such as learning a specialised programming language. Even where these systems can present mathematics using traditional notation, there is usually some non-standard notation or technique for entering expressions to be learnt, and the conventions for such have not become genericised to the same extent as the techniques used by writing tools. The payoff for this effort is that, unlike an “equation editor”, once a mathematical expression has been encoded, the software can transform, solve, manipulate, evaluate and visualise it, relieving the user of many routine tasks.

Unless pupils are expected to become fluent in the use of authentic mathematical tools, to the extent that they (or their conventions and notations) can be called upon in formal assessments, it is unlikely that the computer will become a “natural medium for doing mathematics”. This could mean that the use of computers in mathematics assessment remains, for pupils, a hurdle rather than an advantage.

⁵⁴ <http://demonstrations.wolfram.com/IntersectingLinesUsingSlopeInterceptForm/>

References

- Akker, J. (2006). *Educational design research*. London ;;New York: Routledge.
- AQA. (2003a, June). GCSE Mathematics A Intermediate Paper 2 - June 2003. Assessment and Qualifications Alliance, Leeds, UK.
- AQA. (2003b, November). GCSE Mathematics A Intermediate Paper 1 - November 2003. Assessment and Qualifications Alliance, Leeds, UK.
- AQA. (2006, June). GCSE Mathematics A Intermediate Paper 1 - June 2006. Assessment and Qualifications Alliance, Leeds, UK. Retrieved from <http://web.aqa.org.uk/qual/gcse/qp-ms/AQA-330111-W-QP-JUN06.PDF>
- Babbage, C., & Campbell-Kelly, M. (1994). *Passages from the life of a philosopher*. London: William Pickering.
- Balanced Assessment. (1999). *Balanced Assessment for the Mathematics Curriculum - High School Assessment 1*. White Plains, NY, USA: Dale Seymour.
- Balanced Assessment Project. (1999). *High school assessment package : Berkeley, Harvard, Michigan State, Shell Centre*. Balanced Assessment for the Mathematics Curriculum (Vol. 1). Parsippany NJ: Dale Seymour Publications.
- Barnes, M., Clarke, D., & Stephens, M. (2000). Assessment: the engine of systemic curricular reform? *Journal of curriculum studies*, 32(5), 623–650.

References

- Bell, A. (2003). Domain Frameworks in Mathematics and Problem Solving. Retrieved August 16, 2009, from <http://www.nottingham.ac.uk/education/MARS>
- Black, P. (2008). Strategic Decisions: Ambitions, Feasibility and Context. 1, 1(1). Retrieved from <http://www.educationaldesigner.org/ed/volume1/issue1/article1>
- Black, P., & Wiliam, D. (1998). *Inside the Black Box: Raising Standards Through Classroom Assessment*. King's College, London.
- Bodin, A. (1993). What does To Assess mean? The case of assessing mathematical knowledge. In M. Niss (Ed.), *Investigations into Assessment in Mathematics Education* (pp. 113-141). Dordrecht, NL: Kluwer.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Lawrence Erlbaum Associates Inc.
- Boston, K. (2004). Delivering e-assessment - a fair deal for learners. Speech, . Retrieved from <http://www.qca.org.uk/6998.html>
- Burkhardt, H., & Bell, A. (2007). Problem solving in the United Kingdom. *ZDM*, 39(5), 395-403. doi:10.1007/s11858-007-0041-4
- Burkhardt, H., & Pead, D. (2003). Computer-based assessment: a platform for better tests? In C. Richardson (Ed.), *Whither assessment? : discussions following a seminar, London, March 2002* (pp. 133-148). London: Qualifications and Curriculum Authority.
- Butler,, D., & Hatsell, M. (2007). Autograph. Autograph, Oundle, UK. Retrieved from <http://www.autograph-maths.com/>
- Cabri II Plus. (2009). . Retrieved from <http://www.cabri.com/cabri-2-plus.html>
- California. (2008). Standardized Testing and Reporting (STAR) Results. Retrieved from <http://star.cde.ca.gov/>
- Clausen-May, T., Vappula, H., & Ruddock, G. (2004a). *Progress in Maths 6: Teachers Guide*. London, UK: GL Assessment.
- Clausen-May, T., Vappula, H., & Ruddock, G. (2004b). *Progress in Maths 6: Pupil booklets*. London, UK: GL Assessment.
- Clausen-May, T., Vappula, H., & Ruddock, G. (2004c). *Progress in Maths 7: Teachers Guide*. London,

References

- UK: GL Assessment.
- Cockroft, W. H. (1982). *Mathematics Counts*. London: HMSO.
- Crook, C. (2008). Web 2.0 Activities in Secondary School. Retrieved from <http://jcal.info/web2/>
- Fletcher, L. R. (2009). *MathAssess Final Report*. JISC. Retrieved from <http://www.jisc.ac.uk/Home/publications/documents/mathssassesreport.aspx>
- Geogebra. (2009). . Retrieved from <http://www.geogebra.org/cms/>
- Intelligent Assessment. (2006). Short-Answer Questions on Questionmark Perception. Intelligent Assessment. Retrieved from <http://www.intelligentassessment.com/caseStudy4.htm>
- Johnson, D. C. (1979). Teaching Estimation and Reasonableness of Results. *Arithmetic Teacher*, 27(1), 34-35.
- Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4), 389–405.
- MARS. (2000). *Balanced Assessment in Mathematics (tests)*. Monterey, CA: CTB McGraw-Hill. Retrieved from http://www.ctb.com/mktg/balanced_math/overview.jsp
- Mathematica. (2009). . Retrieved August 16, 2009, from <http://www.wolfram.com/>
- Maxima, a Computer Algebra System. (2008). . Text, . Retrieved August 17, 2009, from <http://maxima.sourceforge.net/>
- Murphy, R., & Noyes, A. (2008). Examination Pilots: where can they take us? Presented at the AEA-Europe Conference, Bulgaria. Retrieved from http://www.aea-europe.net/userfiles/12_Roger%20Murphy_updated.pdf
- Murphy, R., & Noyes, A. (2010). Evaluating Mathematics Pathways (EMP). Retrieved February 22, 2010, from <http://www.nottingham.ac.uk/emp/>
- NCTM. (2000). *Principles and Standards for School Mathematics*. Reston, VA, USA: National Council of Teachers of Mathematics, Inc. Retrieved from www.nctm.org
- Ofsted. (2008). *Mathematics – Understanding the Score*. London: Office for Standards in Education (Ofsted). Retrieved from <http://www.ofsted.gov.uk/>
- Pead, D. (1995). Coypu: A versatile function and data plotter. Shell Centre for Mathematical Education.

References

- Pead, D. (2006). *Progress in Maths 6-14: Analysis and Evaluation of Digital Assessments* (Research report). Mathematics Assessment Resource Service.
- PISA. (2003). *The PISA 2003 assessment framework : mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.
- QCA Curriculum Division. (2007). National Curriculum for England 2008. Retrieved August 14, 2009, from <http://curriculum.qcda.gov.uk/>
- Ridgway, J., Nicholson, J., & McCusker, S. (2006). Reasoning with evidence—New opportunities in assessment. In *Proceedings of the Seventh International Conference on Teaching Statistics, Salvador, Brazil. Voorburg: The Netherlands: International Statistical Institute* . Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/17/6D1_CALL.pdf
- Russel, M. (1999). Testing on Computers: A Follow-up Study Comparing Performance on Computer and On Paper. *Education Policy Analysis Archives* , 7(20). Retrieved from <http://epaa.asu.edu/epaa/v7n20/>
- Sandene, B., Horkay, N., Elliot Bennet, R., Allen, N., Brasswell, J., Kaplan, B., & Oranje, A. (2005). *Online Assessment in Mathematics and Writing* . Washington DC, USA: US Department of Education. Retrieved from <http://nces.ed.gov/nationsreportcard/pdf/studies/2005457.pdf>
- Santos-Bernard, D. (1997). *The Use of Illustrations in School Mathematics Textbooks: Presentation of Information*. The University of Nottingham.
- Schoenfeld, A. (2009). Bridging the Cultures of Educational Research and Design. *Educational Designer*, 1, (2). Retrieved from <http://www.educationaldesigner.org/ed/volume1/issue2/article5/index.htm>
- Schunn, C. (2008). Engineering Educational Design. *Educational Designer*, 1, 1(1). Retrieved from <http://www.educationaldesigner.org/ed/volume1/issue1/article2>
- Schwartz, J. L., Yerushalmy, M., & Wilson, B. (1993). *The Geometric supposer*. Lawrence Erlbaum Associates.
- Shepard, L. A. (1989). Why We Need Better Assessments. *Educational Leadership*, 46(7), 4-9.
- Smith, A. (2004). *Making Mathematics Count* . Department for Education and Science. Retrieved from <http://www.mathsinquiry.org.uk/report/index.html>

References

- Steen, L. (2000). The Case for Quantitative Literacy. In L. Steen (Ed.), *Mathematics and Democracy* (pp. 1-22). National Council on Education and the Disciplines.
- Steen, L. A., & Forman, S. L. (2000). Making Authentic Mathematics Work for All Students. In A. Bessot & J. Ridgway (Eds.), *Education for Mathematics in the Workplace* (pp. 115-126). Kluwer. Retrieved from <http://www.stolaf.edu/people/steen/Papers/authentic.html>
- Swan, M., Pead, D., Crust, R., & Burkhardt, H. (2008). Bowland Maths Professional development resources. Bowland Trust. Retrieved from <http://www.bowlandmaths.org.uk>
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4), 295–312.
- Treilibs, V., Lowe, B., & Burkhardt, H. (1980). *Formulation processes in mathematical modelling*. Nottingham, UK: Shell Centre for Mathematical Education. Retrieved from <http://www.mathshell.com/scp>

Appendix A: A prototype online testing system

This appendix provides some more detail on the design and features of the prototype online testing system developed for the study in Chapter 5.

Design goals

For this study, the key requirements of the system were:

1. The ability to rapidly assemble questions from standard components
2. The flexibility to implement the tools discussed above, and possibly others
3. The ability to capture responses in a flexible format (probably XML)
4. The ability to reconstruct the screen, as seen by the student, from the XML data (for human markers and also to allow students to backtrack and review their answers).
5. Easy installation for trial schools
6. Data to be saved over the internet to a central server as the test progresses. This is based on experiences with WCT which showed that writing data to local storage and subsequently collecting it was a major headache. As this is just a study, we have the luxury of being able to require a broadband internet connection
7. Some degree of crash resilience, such as the ability to resume an interrupted test, possibly on a different machine.
8. Possible use of open-source products, especially for the potentially expensive server-side applications.
9. Cross platform compatibility (PC/Mac/Linux) desirable but not essential.
10. Proof-of-concept that the system could be made accessible to students with special needs if necessary.

The solution adopted uses Adobe Flash for the “client” side and the tasks themselves, since this produces small, internet-friendly applets that resize smoothly to fit the available screen, has a rich programming language, and can be delivered either via a web browser plug-in or as a stand-alone PC or Mac executable. Although we currently rely on the commercial Flash MX 2004 authoring system and (freely distributable) player, there are several potential open-source routes for developing and playing Flash content.

However, Adobe's matching “server-side” products for Flash are fairly expensive commercial products, so the server functions are implemented using an open source webserver (*Apache*), relational database (*PostgreSQL*) and scripting language (*PHP 5*).

The client is compatible with Mac OS X, Windows 2000/XP (and, potentially, Linux). The server is intended to run on a Linux system but should work on other systems with minor modification.

Using the system

Student registration

Schools who agree to take part in the trials are allocated a centre ID and password – they then use their standard web browser to log on to the “admin” site in order to register students and download the necessary client software (Figure A.1).

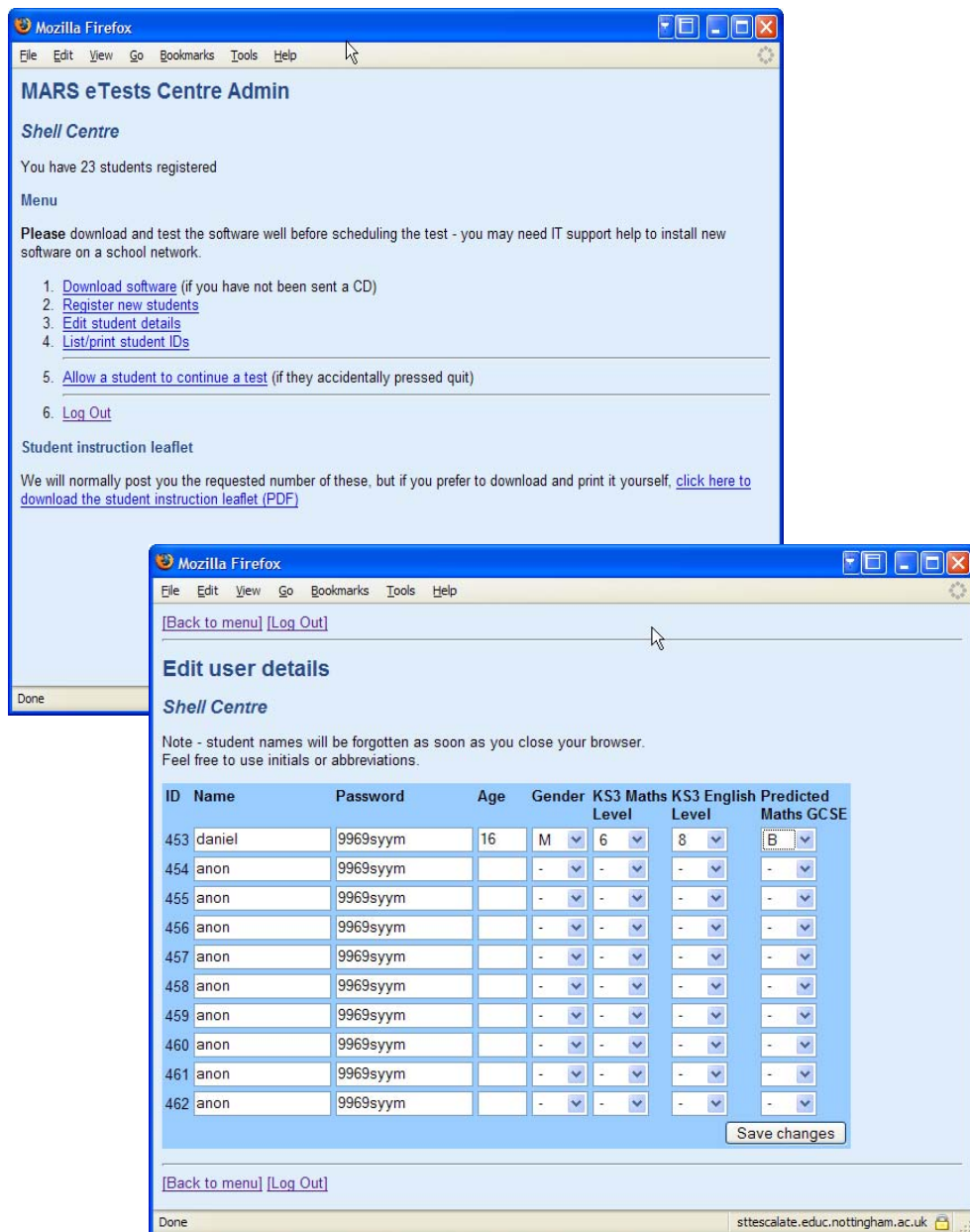


Figure A.1: Student registration system

The trials are “anonymous” – once the teacher has printed out a list of names and ID numbers, the names are wiped and all future correspondence is by ID number. We ask the teacher for age, gender, national test and predicted GCSE performance for each student – it would be very easy to add an ethnicity field for future larger-scale trials. This was omitted from the initial trials since the small number of participant schools was unlikely to produce a representative sample of minority groups.

Download software

The preferred version of the software is a minimal “client” which can be downloaded in a few seconds. Most of the actual software (the test shell and the tasks) is fetched over the internet as and when needed, and the Flash player is incorporated in the client. Consequently, no installation is needed (the client could even be run directly from a floppy disc if needed). Schools were offered various alternatives – including “installers” (which might be needed for some network systems) and versions that included local copies of the test shell and tasks (which put less demand on the internet connection).

All of the variants depend on an internet connection for user authentication and for storage of responses. Since all responses are immediately sent back to the central server, the software does not need to save data to a local hard drive and there is no need for schools to collect and return data – two issues that proved a major headache in the World Class Arena trials.

Log on and take the test

The student starts the software, and logs on using their centre ID, user ID and password. They are then presented with a list of available tests (Figure A.2).

Depending on the permissions granted to the user, they can re-take tests, continue tests they have previously started or review a past test session. In a typical “simulated exam” the student will only be able to take the test once, and then review (but not change) their results.

The tests are presented as a continuous series of pages – individual tasks vary from 1-4 pages in length. Students can move back and forward freely to check and change answers, until they “quit” the test. The trial tests finished with a request for comments (Figure A.3).

The student’s response is sent back to the server as soon as they move on to a new task. If a student changes a response, the server keeps a record of their previous answer(s). Since this system is primarily intended for trials of individual tasks, no time limit or on-screen timer is implemented. However, extensive information on when students started and finished a test and when each task was visited and answered is kept on the server.

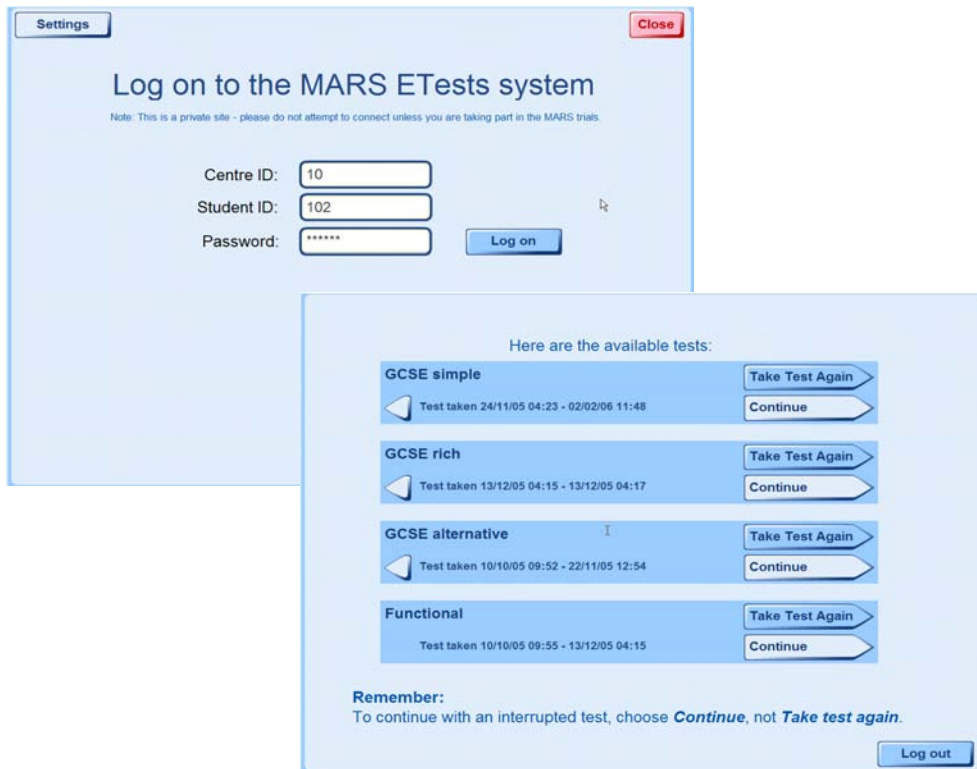


Figure A.2: Logging in

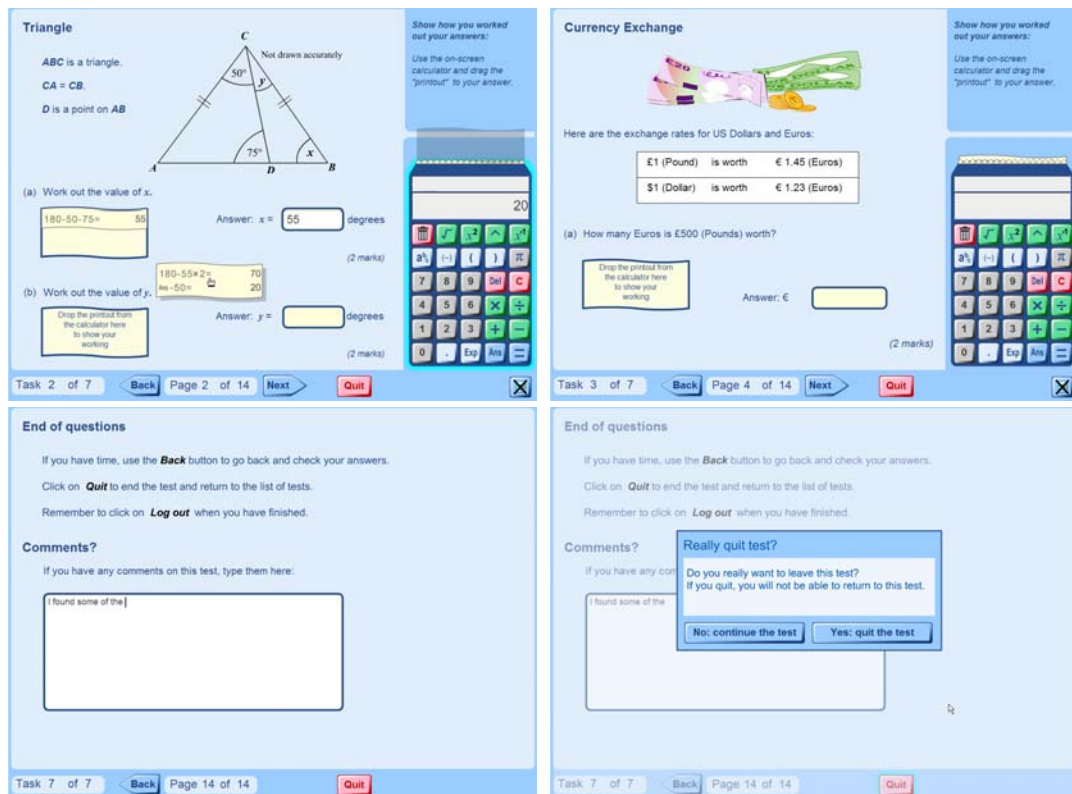


Figure A.3: Taking a test

Marking

Manual marking was performed online, using a system which displayed the mark schemes alongside each student's responses (Figure A.4).

The screenshot shows the Etests Marking system interface. On the left, a list of students is shown with columns for Student ID, Started time, and Tries Mkd?. The student with ID 153 is highlighted in green. The main area displays the marking details for this student, including the task name '4. Trip' and the attempt number 'Attempt 2 of 2'. A table shows the marking scheme and the student's response for each question. The table has columns for Marking, Markscheme, Response, C (Computer mark), and M (Human mark). The responses include text, a graph, and calculator output. The graph shows Distance from Nottingham (miles) on the y-axis (0 to 9) and Time on the x-axis (09:00 to 13:00). The graph shows a piecewise linear function with a peak at 11:00. The calculator output shows '6+2=' and '3'. The human marker has entered '1' in the M column for the first three questions and '0' for the last two.

Marking	Markscheme	Response	C	M
1	Stopped/not moving/taking a rest/stays same distance	B1 Gavin has stopped walking	1	1
2	6 (miles)	B1 Distance= 6 miles	1	1
3.1	(their 6)/2	M1 6+2= 3	1	1
3.2.3	(mph)	A1 Speed = 3 mph	1	1
4	Straight line joining (10:00,9) to (11:00,0)	B1	0	0
5	Time=10:30; Distance=4.5 (miles)	B1 Time = 1130 Distance = 4.5	0	0

Figure A.4: The marking system

The left hand column shows all the “scripts” being marked – “scripts” that have already been marked are highlighted in green. The student’s responses are shown alongside the marking rubric. Graphs and calculator output are reproduced faithfully. Here the question has been marked automatically – the computer’s marks are shown in the “C” column. The human marker can accept the automatic mark for each item by pressing the return key, or alter the mark if they disagree. For the first marking of the trials, the computer marks were hidden and the human markers keyed in their own marks.

Where students have had several attempts at a question, the system tries to select the most extensive/highest scoring answer by default – the marker can review the alternative responses.

The marking notation (“A”, “M”, “B” etc.) is similar to that used in the AQA GCSE.

The system can handle multiple markings of the same responses. The same interface can handle the input of paper test marks. Another screen (Figure A.5) allows responses to be bulk-marked, and compares the results with human markers and previous versions of the marking algorithms.

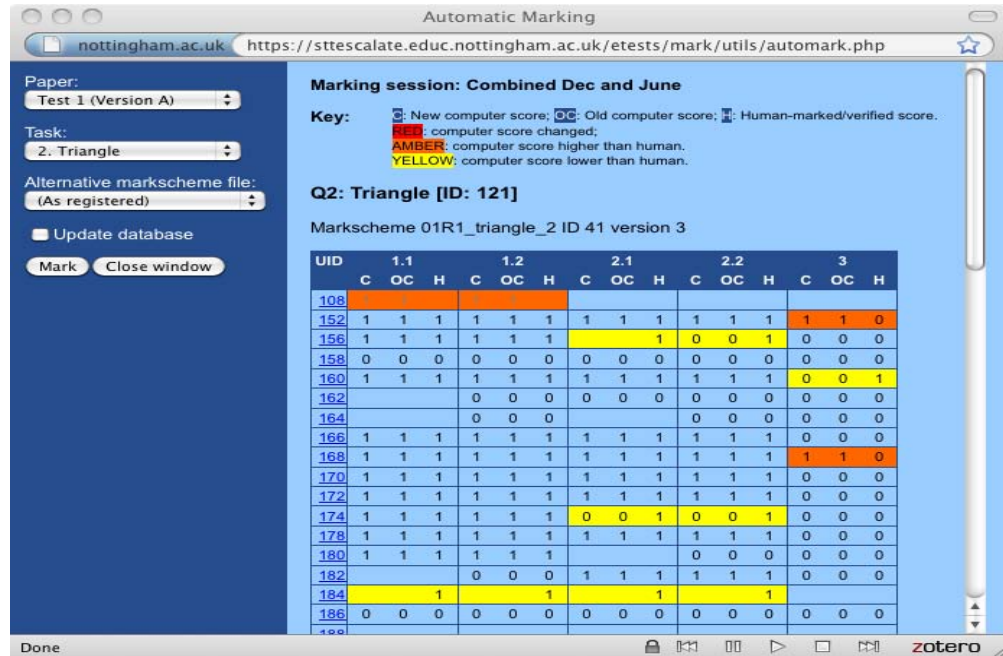


Figure A.5: Testing automatic marking

Results

The task statistics facilities allow a variety of frequency charts, box plots and descriptive statistics to be generated. The data can also be filtered using the demographic data supplied by teachers. In Figure A.6 only pupils at Key Stage 3 levels 5-6 are shown.



Figure A.6: Task statistics display

Behind the scenes

Task design components/tools

Each task is created using the Adobe Flash authoring system (Figure A.7). A number of customised templates, scripts and library items have been produced to facilitate this. Creating a new task is largely a matter of dragging in the required components from the library and setting their parameters (e.g. the grid size and scale labels for a graph). No programming is required, unless a new response/tool type is involved.

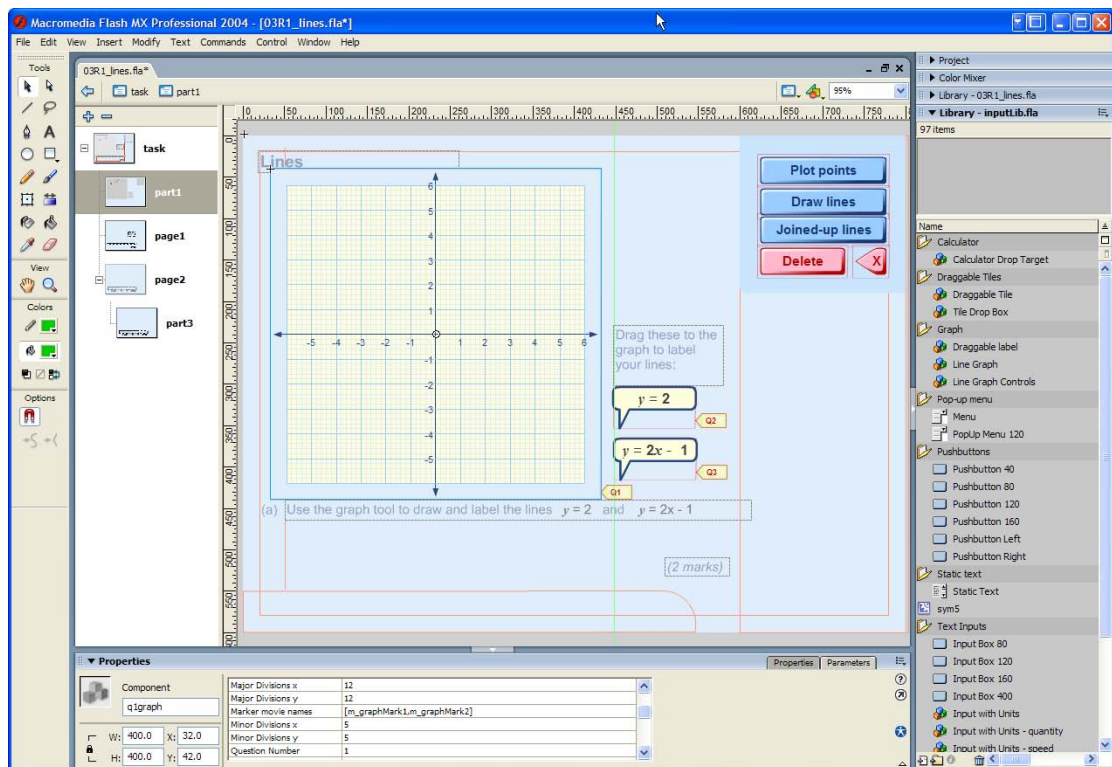


Figure A.7: Creating a task in Adobe Flash

Currently, the library of tool/response “widgets” comprises:

- **Simple text/number inputs** for short answers – these can be configured to allow any text, or to restrict inputs to a particular list of characters or valid decimal/integer numbers
- **Multiple-line text inputs** for longer typed answers.
- **Pushbuttons** which can be arranged into “radio button” groups for multiple-choice responses
- **Pop-up menus** provide another mechanism for multiple choice responses.
- **Calculator drop** (accept “printouts” from the calculator tool)

- **Line Graphs** display a grid & axes and allow the user to draw straight lines and plot points. A graph can also include some pre-defined data (such as a set of points to draw a line through)
- **Draggable labels** carry text or graphics and allow the user to accurately identify points on a graph
- **Draggable tiles** allow for other drag & drop style responses

Mark schemes and marking

A simple “language” – based on the XML mark-up language – was devised to encode mark schemes. The markscheme for each task specifies the following:

- How the students’ responses are to be presented to the markers.
- The “rubric” text used by the human markers to award each point.
- The value and type of each point (e.g. “A1” for an accuracy/final answer mark, “B1” for bonus/independent mark, “M1” for a method mark).#
- The rules for automatic marking. The system includes a series of built-in rules for basic text/numeric value matching (e.g. `<matchvalue min="1.4" max="1.6">` accepts a number between 1.4 and 1.6) plus add-in modules for marking rich answers (the screenshot below uses the linegraph module to match graphical responses based on gradient, proximity to reference points etc.)

One useful feature is that the mark scheme can cross-reference and re-order the responses to individual parts within the task. This is useful if (e.g.) partial credit depends on the answer to previous parts or if method marks need to be assumed if the final answer is correct. It also means that the presentation of a task to the student is less constrained by the logical structure of the mark scheme.

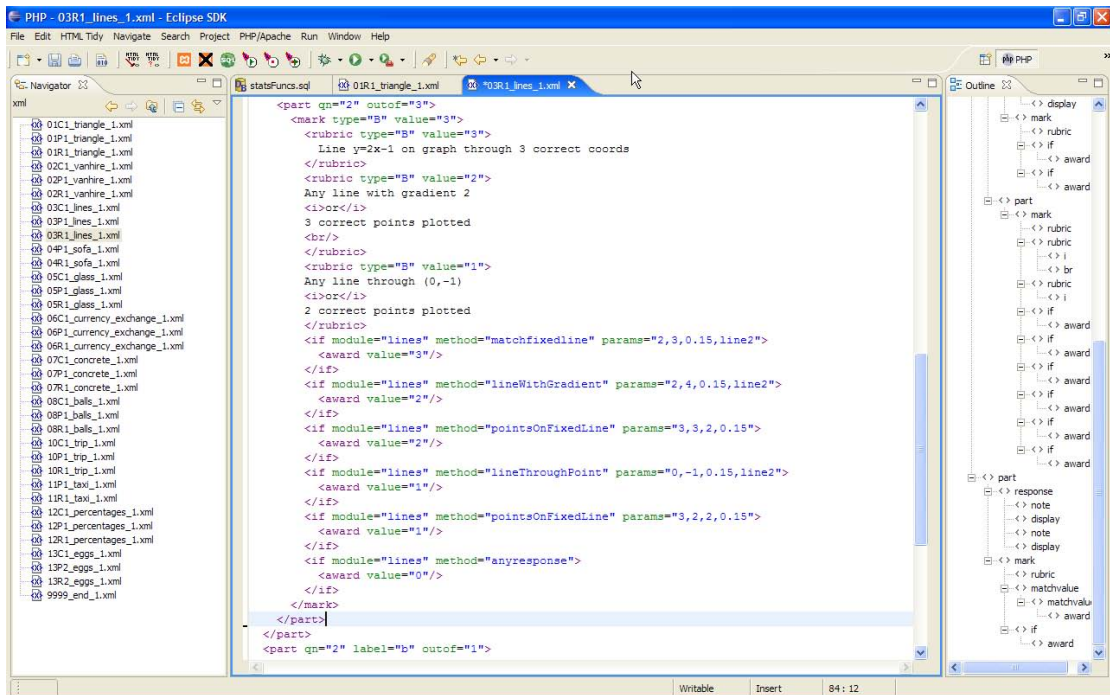


Figure A.8: Creating a mark scheme in XML

Although writing a markscheme by hand is a fairly complex task, XML is designed to be easily generated and interpreted by computer. If the system was to be more widely used then a user-friendly “front end” for creating mark schemes could be created.

The test shell

This is the application that ultimately runs on the user’s machine. The shell is responsible for logging in the user, presenting a menu of tests, loading the task files as required and communicating the responses back to the server.

The shell is written in Flash, and can either be run online via a web browser (provided the Flash Player plug-in is installed) or as a stand-alone application with the appropriate Flash player built-in.

The shell can fetch tasks over the internet (the preferred method, where an adequate internet connection is available) or, optionally (if the stand-alone version is used), from the local hard drive.

Database

An SQL database stores test definitions, candidate login details, responses and results on the central server. This is mostly invisible to users and markers.

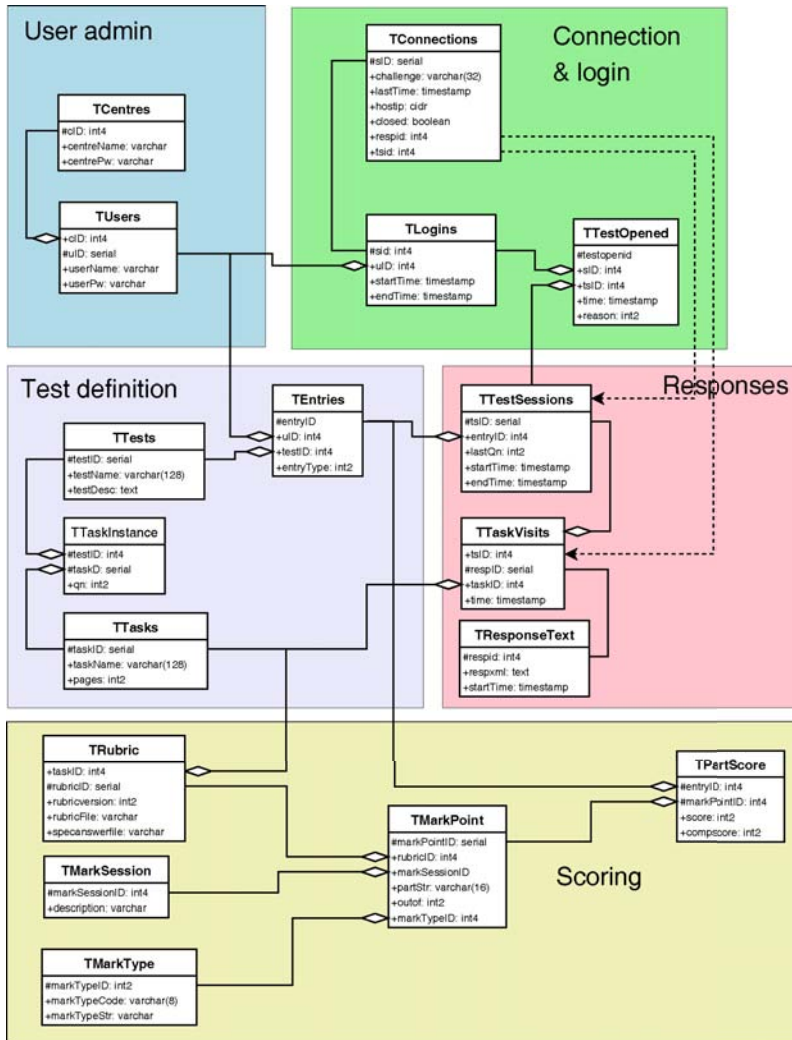


Figure A.9: Schema for the database

Server scripts

Aside from the test client and shell, most of the “logic” behind the system is handled by scripts (written in the PHP language) running on the server.

The test delivery scripts provided an XML-based communication interface between the shells running on client computers and the database.

Other scripts manage a series of interactive web pages – accessed using a standard web browser – that handle student registration, marking and results presentation.

Scripts can also be written which add automatic marking rules to the markscheme “language” in order to accommodate new response types or more sophisticated analysis of existing types.

Accessibility

Although accessibility by students with special needs was not an issue during the trials, future use of the system could require this, so some of the basic requirements of accessibility have been addressed:

- Everything - including the calculator and graph tools - can be driven from the keyboard and hence, in principle, by switch devices, concept keyboards etc. Drag and drop operations can be keyboard driven (“space” to pick up, cursor keys to move, “space” to drop).
- The principle colours of all buttons and text can be changed (currently there is just the regular colour scheme and a black and white scheme – more can be added).
- Where graphics are used, it is possible to incorporate several versions of the graphic with different colour schemes.
- Flash applications will scale to fit a large screen and have a zoom facility.
- It should be possible to incorporate screen reader support (although this would need extensive testing to ensure that it was a feasible alternative).

These facilities have not been fully utilised or tested in the current system – the object was to ensure that they could be incorporated (at short notice) if needed. Extensive testing would be needed before the system could be advertised as “Accessible”.

However, there are wider and deeper questions over how assessment (particularly some of the aspects unique to mathematics) can be made truly accessible, which is beyond the scope of this study.

Appendix B: Materials in electronic form

To save space in the printed document, further background information on the work, including some live software, is supplied in electronic form on the accompanying CD or online at: <http://www.mathshell.org/papers/dpthesis/>

Selection of tasks from World Class Tests

Working versions of the tasks discussed in Chapter 3.

Report to nferNelson on Progress in Maths (Pead, 2006)

This includes the full task-by-task analysis and discussion of the tasks in the tests reviewed in Chapter 4 (not available online).

The complete task set used in Chapter 5, comprising:

- The paper tasks
- Screen shots of the computer based versions
- Mark schemes
- Working previews of the tasks